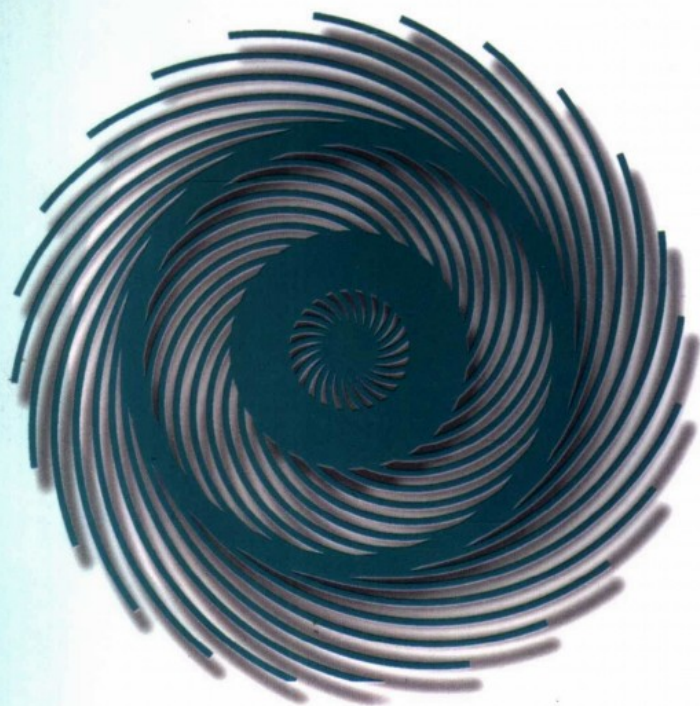
颜庆津　编著



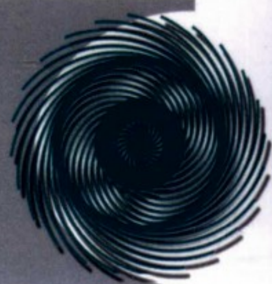
高等学校研究生教材

数值分析（第3版）



北京航空航天大学出版社

GAODENG XUEXIAO
YANJIUSHENG
JIAOCAI



策　划：胡　敏
责任编辑：宋淑娟
书籍装帧：



ISBN 7-81077-877-3
9 787810 778770 >



ISBN 7-81077-877-3
定价：22.00元

高等学校研究生教材

数值分析

（第3版）

颜庆津　编著

北京航空航天大学出版社

内 容 简 介

本书是为工学硕士研究生数值分析课而编写的学位课教材。内容包括：线性方程组的解法，矩阵特征值与特征向量的计算，非线性方程与非线性方程组的迭代解法，插值与逼近，数值积分，常微分方程初值问题的数值解法和偏微分方程的差分解法。内容丰富，系统性强，语言简练、流畅，数值例子和习题非常丰富，并附习题答案。其深度和广度适合工学硕士生的培养要求。

本书还可供从事科学与工程计算的科技人员自学和参考。

图书在版编目(CIP)数据

数值分析/颜庆津编著. —3版. —北京：北京航空航天大学出版社，2006.7

ISBN 7-81077-877-3

Ⅰ.数… Ⅱ.颜… Ⅲ.数值计算—研究生—教材
Ⅳ.O241

中国版本图书馆CIP数据核字（2006）第050782号

数值分析(第3版)
颜庆津　编著
责任编辑　宋淑娟
*
北京航空航天大学出版社出版发行
北京市海淀区学院路37号(100083)　发行部电话(010)82317024　传真(010)82328026
http://www.buaapress.com.cn　E-mail:bhpress@263.net
北京市松源印刷有限公司印装　各地书店经销
*
开本：787×1092　1/16　印张：16.75　字数：429千字
2006年7月第3版　2006年7月第1次印刷　印数：5 000册
ISBN 7-81077-877-3　定价：22.00元

第 3 版序

本书是工学硕士研究生数值分析课的基本教材，是作者继 1992 年出版的第 1 版和 2000 年出版的修订版之后编著的数值分析第 3 版。2000 年出版的《数值分析(修订版)》经过多所院校六年的教学实践，证明该书确实贯彻了重概念、重方法、重应用、重能力培养的原则，其内容的深度和广度确实符合工学硕士研究生的培养要求，得到了广大使用者的欢迎和肯定。

本书与修订版相比，主要是在教学法上做了较大的改进，体现了几年来使用修订版的一些成功的教学经验；习题也做了相应的调整，并在书的最后增加了全书习题的答案与提示；此外，还改正了修订版中的疏漏之处，使全书的叙述更加严谨。

作　者

2006.4

修订版前言

本书是为工学硕士研究生数值分析课而编写的学位课教材，是在作者 1992 年编写的《数值分析》(北京航空航天大学出版社，1992.7)的基础上修订而成的。它仍然遵循重概念、重方法、重应用、重能力培养的原则，并针对工学硕士研究生的培养要求，使学生掌握一定的理论深度。

与第 1 版相比，本书在内容的深度和广度上均做了较大的调整。一方面尽量简化在本科计算方法课中已有的内容，减少重复；另一方面新增加了一些目前在科学技术中需要使用的数值方法及其有关理论，使其更适应当前工学硕士研究生的培养需求。

只须具备工科本科高等数学和线性代数的知识，就能学习本书的内容。如果还掌握了一种计算机程序设计语言，并能上机计算实习，则对本书的内容会有更深刻的体会。讲授本书的全部内容大约需要 70 学时。学时数少于 70 的，可对各章内容选择讲授。本书每章都附有习题，使用它作教材的研究生都应以这些习题作为基本练习。

本书出版前，由清华大学数学科学系关治教授审阅了全部书稿，并提出了重要的修改意见，对此深表感谢。

作　者

1999.6

目　　录

第1章　绪　论

第2章　线性方程组的解法

第3章　矩阵特征值与特征向量的计算

第1章 绪 论

1.1 数值分析的研究对象

现代科学技术问题的研究方法可分为三种：理论推导、科学实验和科学计算。这三种方法相辅相成，又相互独立且缺一不可。科学计算就是通过建立数学模型把科学技术问题转化为数学问题，然后对数学问题进行离散化，将其转化为数值问题，最后使用数值计算方法计算出数值问题的解，并把所得的解作为原科学技术问题的解。随着电子计算机的性能不断提高，科学计算在解决现代科学技术问题中所起的作用越来越大，并已渗透到科学技术的各个领域。科学计算的基础——计算数学这个数学分支也随之发展壮大。数值分析是计算数学中最基本的内容。它研究如何用数值计算方法求解各种基本数学问题以及在求解过程中出现的收敛性、数值稳定性和误差估计等问题。数值分析所阐明的各种数值计算方法是从事科学计算的最基本工具。

1.2 误差知识与算法知识

1.2.1 误差的来源与分类

在工程技术的计算中，估计计算结果的精确度是十分重要的工作，而影响精确度的是各种各样的误差。误差按照它们的来源可分为以下四种。

1. 模型误差

反映实际问题有关量之间关系的计算公式，即数学模型，通常只是近似的。由此产生的数学模型的解与实际问题的解之间的误差称为模型误差。

2. 观测误差

数学模型中包含的某些参数（如时间、长度、电压等）往往通过观测而获得。由观测得到的数据与实际的数据之间是有误差的。这种误差称为观测误差。

3. 截断误差

求解数学模型所用的数值计算方法如果是一种近似的方法，那么只能得到数学模型的近似解，由此产生的误差称为截断误差或方法误差。例如，由 Taylor（泰勒）公式，函数 $f(x)$ 可表示为

$$f(x)=f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n+\frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1},\quad 0<\theta<1$$

为了简化计算，当$|x|$不大时，去掉上式右端的最后一项，得近似公式

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n$$

此近似公式的误差就是截断误差。

4. 舍入误差

由于计算机的字长有限，参加运算的数据及其运算结果在计算机上存放会产生误差。这种误差称为舍入误差或计算误差。例如，在十位十进制的限制下，会出现

$$1 \div 3 = 0.333\,333\,333\,3$$

$$(1.000\,002)^2 - 1.000\,004 = 0$$

两个结果都不是准确的，后者的准确结果应是4×10^{-12}。这里所产生的误差就是舍入误差。

在数值分析中，主要研究截断误差和舍入误差对计算结果的影响，而一般不考虑模型误差和观测误差。

1.2.2 绝对误差、相对误差与有效数字

设a是准确值x的一个近似值，记

$$e = x - a$$

称e为近似值a的绝对误差，简称误差。如果$|e|$的一个上界已知，记为ε，即

$$|e| \leqslant \varepsilon$$

则称ε为近似值a的绝对误差限或绝对误差界，简称误差限或误差界。

准确值x、近似值a和误差限ε三者的关系就是

$$a - \varepsilon \leqslant x \leqslant a + \varepsilon$$

或记为

$$x = a \pm \varepsilon$$

例如，$a=3.14$作为圆周率π的一个近似值，它的绝对误差是

$$e = \pi - 3.14$$

易知，

$$|e| < 0.002$$

所以，$a=3.14$作为π的近似值，它的一个误差限为

$$\varepsilon = 0.002$$

用绝对误差来刻画近似值的精确程度是有局限性的，因为它没有反映出其本身在原数中所占的比例。

记

$$e_r = \frac{e}{x} = \frac{x-a}{x}$$

称e_r为近似值a的相对误差。由于x未知，实际上总把$\frac{e}{a}$作为a的相对误差，并且也记为

$$e_r = \frac{e}{a} = \frac{x-a}{a}$$

相对误差一般用百分比表示。

$|e_r|$的上界，即

$$\varepsilon_r = \frac{\varepsilon}{|a|}$$

称为近似值 a 的相对误差限或相对误差界。显然有

$$|e_r| \leqslant \varepsilon_r$$

例 1　用最小刻度为毫米的卡尺测量直杆甲和直杆乙，分别读出长度 $a=312$ mm 和 $b=24$ mm，问：$\varepsilon(a)$，$\varepsilon(b)$，$\varepsilon_r(a)$，$\varepsilon_r(b)$各是多少？两直杆实际长度 x 和 y 在什么范围内？

解

$$\varepsilon(a)=\varepsilon(b)=0.5\ \text{mm}$$

$$\varepsilon_r(a) = \frac{\varepsilon(a)}{|a|} = \frac{0.5}{312} \approx 0.16\%$$

$$\varepsilon_r(b) = \frac{\varepsilon(b)}{|b|} = \frac{0.5}{24} \approx 2.08\%$$

$$311.5\ \text{mm} \leqslant x \leqslant 312.5\ \text{mm}$$

$$23.5\ \text{mm} \leqslant y \leqslant 24.5\ \text{mm}$$

例 2　设 $a=-2.18$ 和 $b=2.120\,0$ 是分别由准确值 x 和 y 经过四舍五入而得到的近似值，问：$\varepsilon(a)$，$\varepsilon(b)$，$\varepsilon_r(a)$，$\varepsilon_r(b)$各是多少？

解

$$\varepsilon(a)=0.005,\quad \varepsilon(b)=0.000\,05$$

$$\varepsilon_r(a) = \frac{0.005}{2.18} \approx 0.23\%$$

$$\varepsilon_r(b) = \frac{0.000\,05}{2.120\,0} \approx 0.002\,4\%$$

凡是由准确值经过四舍五入而得到的近似值，其绝对误差限等于该近似值末位的半个单位。

定义　设数 a 是数 x 的近似值。如果 a 的绝对误差限是它的某一位的半个单位，并且从该位到它的第一位非零数字共有 n 位，则称用 a 近似 x 时具有 n 位有效数字。

非零小数 a 总可以写成如下的形式

$$a = \pm 0.a_1a_2\cdots a_k \times 10^m$$

其中 m 是整数，$a_i(i=1,2,\cdots,k)$是 0 到 9 中的一个数字，$a_1 \neq 0$。如果 a 作为数 x 的近似值，且

$$\varepsilon(a) = \frac{1}{2} \times 10^{m-n},\quad n \leqslant k$$

则由定义知，a 有 n 位有效数字 $a_1,a_2,\cdots,a_n$。

从有效数字的定义可知，由准确值经过四舍五入得到的近似值，从它的末位数字到第一位非零数字都是有效数字。同一个准确值的不同近似值，有效数字越多，其绝对误差和相对误差都越小。有了有效数字概念之后，下面 2.140 012 的两个近似值 2.14 和 2.140 0 的写法是有区别的。前者有三位有效数字，后者有五位有效数字。

准确值的有效数字可看做有无限多位。

例 3　下列近似值的绝对误差限都是 0.005，

$$a = 1.38,\quad b = -0.031\,2,\quad c = 0.86 \times 10^{-4}$$

问：各个近似值有几位有效数字？

解 a 有三位有效数字 1,3,8。b 有一位有效数字 3。c 没有有效数字。

1.2.3 函数求值的误差估计

设 $u=f(x)$ 存在足够高阶的导数,a 是自变量 x 的近似值,则 $\tilde{u}=f(a)$ 是函数值 $u=f(x)$ 的近似值。如果 $f'(a)\neq 0$ 且比值 $|f''(a)|/|f'(a)|$ 不很大,则由 Taylor 公式可得 $\tilde{u}=f(a)$ 的误差估计为

$$e(\tilde{u}) = f(x)-f(a) \approx f'(a)(x-a) = f'(a)e(a)$$

因
$$|e(\tilde{u})| \approx |f'(a)|\,|e(a)| \leqslant |f'(a)|\varepsilon(a)$$

故
$$\varepsilon(\tilde{u}) \approx |f'(a)|\,\varepsilon(a)$$

如果 $f'(a)=f''(a)=\cdots=f^{(k-1)}(a)=0$, $f^{(k)}(a)\neq 0$,且比值 $|f^{(k-1)}(a)|/|f^{(k)}(a)|$ 不很大,则 $\tilde{u}=f(a)$ 的误差估计为

$$e(\tilde{u}) \approx \frac{f^{(k)}(a)}{k!}[e(a)]^k$$

$$\varepsilon(\tilde{u}) \approx \frac{|f^{(k)}(a)|}{k!}[\varepsilon(a)]^k$$

设 n 元函数 $u=f(x_1,x_2,\cdots,x_n)$ 充分可微,a_i 是 x_i 的近似值,其中 $i=1,2,\cdots,n$,则 $\tilde{u}=f(a_1,a_2,\cdots,a_n)$ 是函数值 $u=f(x_1,x_2,\cdots,x_n)$ 的近似值。由多元函数 Taylor 公式可得 $\tilde{u}$ 的误差估计为

$$e(\tilde{u}) \approx \sum_{i=1}^{n} \frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i} e(a_i)$$

$$\varepsilon(\tilde{u}) \approx \sum_{i=1}^{n} \left| \frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i} \right| \varepsilon(a_i) \tag{1.1}$$

如果 $\left| \frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i} \right|$ 全为零或全都很小,则要使用 Taylor 公式中的高阶项。

由式(1.1)可推出四则运算结果的误差估计。设 a 和 b 分别是准确值 x 和 y 的近似值,则 $a+b$, $a-b$, ab, $a/b(b\neq 0)$ 分别是 $x+y$, $x-y$, xy, x/y 的近似值。根据式(1.1),可得

$$\varepsilon(a\pm b) = \varepsilon(a)+\varepsilon(b)$$

$$\varepsilon(ab) \approx |a|\,\varepsilon(b)+|b|\,\varepsilon(a)$$

$$\varepsilon\left(\frac{a}{b}\right) \approx \frac{|a|\,\varepsilon(b)+|b|\,\varepsilon(a)}{|b|^2}, \quad b\neq 0$$

$$\varepsilon_r(a+b) = \frac{\varepsilon(a)+\varepsilon(b)}{|a+b|}$$

$$\varepsilon_r(a-b) = \frac{\varepsilon(a)+\varepsilon(b)}{|a-b|}$$

$$\varepsilon_r(ab) \approx \varepsilon_r(a)+\varepsilon_r(b)$$

$$\varepsilon_r\left(\frac{a}{b}\right) \approx \varepsilon_r(a)+\varepsilon_r(b)$$

例 4 设有三个近似数

$$a=2.31, \quad b=1.93, \quad c=2.24$$

它们都有三位有效数字,试计算 $p=a+bc$, $\varepsilon(p)$ 和 $\varepsilon_r(p)$,并问:p 的计算结果能有几位有效数字?

解
$$p=2.31+1.93\times2.24=6.633\ 2$$
$$\begin{aligned}\varepsilon(p)&=\varepsilon(a)+\varepsilon(bc)\approx\\&\varepsilon(a)+|b|\varepsilon(c)+|c|\varepsilon(b)=\\&0.005+0.005(1.93+2.24)=0.025\ 85\end{aligned}$$
$$\varepsilon_r(p)=\frac{\varepsilon(p)}{|p|}\approx\frac{0.025\ 85}{6.633\ 2}\approx0.39\%$$

因为 $\varepsilon(p)\approx0.025\ 85<0.05$，所以 $p=6.633\ 2$ 中能有两位有效数字。

例 5 设 $f(x,y)=\frac{\cos y}{x}$，$x=1.30\pm0.005$，$y=0.871\pm0.000\ 5$。如果用 $\tilde{u}=f(1.30,0.871)$ 作为 $f(x,y)$ 的近似值，则 $\tilde{u}$ 能有几位有效数字？

解
$$\tilde{u}=\frac{\cos 0.871}{1.30}\approx0.495\ 43$$

由于

$$\frac{\partial f}{\partial x}=-\frac{\cos y}{x^2},\quad \frac{\partial f}{\partial y}=-\frac{\sin y}{x}$$

所以

$$\varepsilon(\tilde{u})\approx\left|\frac{\cos 0.871}{1.30^2}\right|\times0.005+\left|\frac{\sin 0.871}{1.30}\right|\times0.000\ 5\approx0.002\ 2<0.005$$

因而 $\tilde{u}=0.495\ 43$ 能有两位有效数字。

1.2.4 算法及其计算复杂性

用数值计算方法求解数值问题是通过具体的算法实现的。所谓算法就是规定了怎样从输入数据计算出数值问题解的一个有限的基本运算序列。其中，基本运算是指四则运算、逻辑运算和一些基本函数运算。衡量算法的优劣有两个标准：其一是要有可靠的理论基础，包括正确性、收敛性、数值稳定性以及可作误差分析；其二是要有良好的计算复杂性。

算法的计算复杂性是指在达到给定精度时该算法所需的计算量和所占的内存空间。前者称为时间复杂性，后者称为空间复杂性。在同一精度要求下，算法所需的计算量少，称为时间复杂性好；所占的内存空间少，称为空间复杂性好。例如，计算多项式

$$p(x)=\sum_{j=0}^{n}a_jx^j$$

的值，输入数据为 $a_j(j=0,1,\cdots,n)$ 和 x，输出数据为 $p(x)$ 值。

算法一

$$\begin{cases}s_0=a_0\\s_k=a_kx^k\quad(k=1,2,\cdots,n)\\p(x)=s_0+s_1+\cdots+s_n\end{cases}$$

算法二（秦九韶法）

$$\begin{cases}T_n=a_n\\T_k=xT_{k+1}+a_k\quad(k=n-1,n-2,\cdots,1,0)\\p(x)=T_0\end{cases}$$

算法一所需的乘法次数为 $1+2+\cdots+n=\frac{n(n+1)}{2}$，加法次数为 n；算法二所需的乘法次

数和加法次数都是 n。两种算法所占的内存空间基本相同。可见，算法二的计算复杂性优于算法一。算法二是公元 1247 年我国古代数学家秦九韶在世界上首次提出的。

算法通常是在计算机上执行的，而计算机存储数据的字长有限，因而产生舍入误差。为了减少舍入误差的影响，设计算法时应遵循以下一些原则：

(1) 要有数值稳定性，即能控制舍入误差的传播。例如，要在四位十进制的限制下计算积分

$$y_n = \int_0^1 \frac{x^n}{x+5}\mathrm{d}x \quad (n = 0,1,\cdots,100)$$

利用关系式 $y_n + 5y_{n-1} = \frac{1}{n}$，可得出如下的算法：

$$\begin{cases} y_0 = \ln 6 - \ln 5 \approx 0.1823 \\ y_n = \dfrac{1}{n} - 5y_{n-1} \quad (n = 1,2,\cdots,100) \end{cases}$$

这个算法显然不具有数值稳定性，因为 $y_0 \approx 0.1823$ 的舍入误差传给 y_1 时，就增至 5 倍，传到 y_{100} 时将是 5^{100} 倍。现利用估计式

$$\frac{1}{6(n+1)} < y_n < \frac{1}{5(n+1)}$$

并取 $y_{100} \approx \frac{1}{2}\left(\frac{1}{606} + \frac{1}{505}\right) \approx 0.001815$，得出另一算法

$$\begin{cases} y_{100} \approx 0.001815 \\ y_{n-1} = \dfrac{1}{5n} - \dfrac{1}{5}y_n \quad (n = 100,99,\cdots,1) \end{cases}$$

这个算法就具有数值稳定性。

(2) 两数相加要防止较小的数加不到较大的数中所引起的严重后果。较小的数加不到较大的数中有时是允许的，但有时会产生严重的后果。例如，在十位十进制的限制下求解一元二次方程

$$x^2 + 10^4 x - 0.01 = 0$$

并且使用求根公式

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

这时，按照加法运算的对阶规则，应有

$$b^2 - 4ac = 10^8 + 0.04 = 0.1 \times 10^9 + 0.00000000004 \times 10^9$$

由于是在十位十进制的限制下进行运算，所以，上式中的 $0.00000000004 \times 10^9$ 被当做是 0，因而

$$b^2 - 4ac = 0.1 \times 10^9 = 10^8$$

于是，得到

$$x_1 = 0, \quad x_2 = -10^4$$

所求得的根中，$x_2 = -10^4$ 是合理的，可接受的；但 $x_1 = 0$ 是不可接受的，其后果是严重的。为了避免后一种情况的出现，计算 x_1 时，可利用关系式

$$x_1 x_2 = -0.01$$

由此得

$$x_1 = -\frac{0.01}{-10^4} = 10^{-6}$$

(3) 要尽量避免两个相近的近似值相减，以免严重损失有效数字。例如，$x=1.232$，$y=1.231$是两个准确值。现要在四位十进制的限制下计算 $z=x^3-y^3$ 的值。一种算法是按给出的式子直接计算，得

$$z = 1.232^3 - 1.231^3 \approx 1.870 - 1.865 = 0.005$$

所得结果最多有一位有效数字，原因是出现了 1.870 和 1.865 这两个相近的近似值相减。另外一种算法是

$$z = x^3 - y^3 = (x-y)(x^2+xy+y^2) \approx$$
$$0.001 \times (1.518 + 1.517 + 1.515) = 0.004\,550$$

其中 0.001 是准确值，因而

$$\varepsilon(z) = 0.001 \times (0.000\,5 + 0.000\,5 + 0.000\,5) =$$
$$0.000\,001\,5 < 0.000\,005$$

由此可知，$z \approx 0.004\,550$ 至少有三位有效数字。

(4) 除法运算中，要尽量避免除数的绝对值远远小于被除数的绝对值。当 a,b 中有近似值时，由

$$\varepsilon\left(\frac{a}{b}\right) \approx \frac{|a|\,\varepsilon(b) + |b|\,\varepsilon(a)}{|b|^2}, \quad b \neq 0$$

可知，如果$|b| \ll |a|$，则 $\varepsilon\left(\frac{a}{b}\right)$可能很大。当 a,b 都是准确值时，由于$\left|\frac{a}{b}\right|$很大，会使其他较小的数加不到$\frac{a}{b}$中而引起严重后果。

1.3　向量范数与矩阵范数

1.3.1　向量范数

向量范数是用于定义向量大小的量，又称为向量的模。

定义　定义在 $\mathbf{R}^n$ 上的实值函数 $\|\cdot\|$ 称为向量范数，如果对于 $\mathbf{R}^n$ 中的任意向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，它满足

(1) 正定性：$\|\boldsymbol{x}\| \geqslant 0$，当且仅当 $\boldsymbol{x}=\mathbf{0}$ 时，$\|\boldsymbol{x}\|=0$；

(2) 齐次性：对任一数 $k \in \mathbf{R}$，有 $\|k\boldsymbol{x}\| = |k|\,\|\boldsymbol{x}\|$；

(3) 成立三角不等式：$\|\boldsymbol{x}+\boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$。

定理 1.1　对 $\mathbf{R}^n$ 中的任一向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，记

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\|x\|_\infty = \max_{1\leqslant i\leqslant n} |x_i|$$

则 $\|\cdot\|_1$，$\|\cdot\|_2$ 和 $\|\cdot\|_\infty$ 都是向量范数。

证 只证 $\|\cdot\|_2$ 是向量范数，其余两个留给读者自己证明。

$\|\cdot\|_2$ 满足定义中的条件(1)是显然的。对任一数 $k\in\mathbf{R}$，有

$$\|kx\|_2 = \sqrt{\sum_{i=1}^n (kx_i)^2} = \sqrt{k^2\sum_{i=1}^n x_i^2} = |k|\sqrt{\sum_{i=1}^n x_i^2} = |k|\ \|x\|_2$$

因此，$\|\cdot\|_2$ 满足定义中的条件(2)。由 $\|x\|_2$ 的含义，可用内积表示 $\|x\|_2$，即

$$\|x\|_2 = \sqrt{x^{\mathrm{T}}x}$$

任取向量 $y\in\mathbf{R}^n$，则有

$$\|x+y\|_2^2 = (x+y)^{\mathrm{T}}(x+y) = \|x\|_2^2 + 2x^{\mathrm{T}}y + \|y\|_2^2$$

根据 Cauchy - Schwarz(柯西-施瓦兹)不等式

$$(x^{\mathrm{T}}y)^2 \leqslant (x^{\mathrm{T}}x)(y^{\mathrm{T}}y)$$

可知

$$\|x+y\|_2^2 \leqslant \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 = (\|x\|_2 + \|y\|_2)^2$$

因而 $\|\cdot\|_2$ 满足定义中的条件(3)。

证毕。

称 $\|\cdot\|_1$ 为 1 -范数或列范数；称 $\|\cdot\|_2$ 为 2 -范数或Euclid(欧几里得)范数，$\|x\|_2$ 实际上就是 n 维向量空间中向量 x 的欧氏长度；称 $\|\cdot\|_\infty$ 为∞-范数或行范数。其实，它们都是 p -范数

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$$

的特例，其中，正整数 $p\geqslant 1$，并且有 $\lim\limits_{p\to\infty}\|x\|_p = \max\limits_{1\leqslant i\leqslant n}|x_i|$。

在空间 $\mathbf{R}^n$ 中可以引进各种向量范数，它们都满足下述向量范数等价定理。

定理 1.2 设 $\|\cdot\|_\alpha$，$\|\cdot\|_\beta$ 是 $\mathbf{R}^n$ 上的任意两种向量范数，则存在与向量 x 无关的常数 m 和 $M(0<m<M)$，使下列关系成立

$$m\|x\|_\alpha \leqslant \|x\|_\beta \leqslant M\|x\|_\alpha,\quad \forall x\in\mathbf{R}^n$$

(证明从略)

定理 1.2 的意义在于，向量 x 的某一种范数可以任意小(大)时，该向量的其他任何一种范数也会任意小(大)。

当不需要指明使用哪一种向量范数时，就用记号 $\|\cdot\|$ 泛指任何一种向量范数。

1.3.2 矩阵范数

矩阵范数是用于定义矩阵“大小”的量，又称为矩阵的模。

定义 定义在 $\mathbf{R}^{n\times n}$ 上的实值函数 $\|\cdot\|$ 称为矩阵范数，如果对于 $\mathbf{R}^{n\times n}$ 中的任意矩阵 A 和 B，它满足

(1) $\|A\|\geqslant 0$，当且仅当 $A=O$ 时，$\|A\|=0$；

(2) 对任一数 $k\in\mathbf{R}$，有 $\|kA\|=|k|\ \|A\|$；

(3) $\|A+B\|\leqslant\|A\|+\|B\|$；

(4) $\|AB\| \leqslant \|A\|\ \|B\|$。

在矩阵计算中，矩阵与向量的乘积经常出现，因而应让所用的矩阵范数与向量范数有某种关系。

定义　对于给定的向量范数 $\|\cdot\|$ 和矩阵范数 $\|\cdot\|$，如果对任一个 $x \in \mathbf{R}^n$ 和任一个 $A \in \mathbf{R}^{n\times n}$，满足

$$\|Ax\| \leqslant \|A\|\ \|x\|$$

则称所给的矩阵范数与向量范数是相容的。

当定义一种矩阵范数时，应当使它能与某种向量范数相容。当在同一个问题中需要同时使用矩阵范数和向量范数时，这两种范数应当是相容的。

现在给出一种定义矩阵范数的方法。

定理 1.3　设在 $\mathbf{R}^n$ 中给定了一种向量范数，对任一矩阵 $A \in \mathbf{R}^{n\times n}$，令

$$\|A\| = \max_{\|x\|=1} \|Ax\| \tag{1.2}$$

则由式(1.2)定义的 $\|\cdot\|$ 是一种矩阵范数，并且它与所给定的向量范数相容。

证　首先证明相容性。对任意的矩阵 $A \in \mathbf{R}^{n\times n}$ 和任意的非零向量 $y \in \mathbf{R}^n$，由于

$$\max_{\|x\|=1} \|Ax\| \geqslant \left\| A \frac{y}{\|y\|} \right\| = \frac{1}{\|y\|} \|Ay\|$$

所以有

$$\|Ay\| \leqslant \|y\| \max_{\|x\|=1} \|Ax\| = \|A\|\ \|y\|$$

此结果显然也适用于 $y=\mathbf{0}$ 的情形。

再证明式(1.2)满足矩阵范数的四个条件。

(1) 当 $A=O$ 时，$\|A\|=0$；当 $A \neq O$ 时，必有 $\|A\|>0$。

(2) 对任一数 $k \in \mathbf{R}$，有

$$\|kA\| = \max_{\|x\|=1} \|kAx\| = |k| \max_{\|x\|=1} \|Ax\| = |k|\ \|A\|$$

(3) 对任意的矩阵 $A, B \in \mathbf{R}^{n\times n}$，式

$$\begin{aligned}\|A+B\| &= \max_{\|x\|=1} \|(A+B)x\| = \\ &\max_{\|x\|=1} \|Ax+Bx\| \leqslant \max_{\|x\|=1} (\|Ax\| + \|Bx\|) \leqslant \\ &\max_{\|x\|=1} \|Ax\| + \max_{\|x\|=1} \|Bx\| = \|A\| + \|B\|\end{aligned}$$

成立。

(4) $\|AB\| = \max\limits_{\|x\|=1} \|(AB)x\| \leqslant \max\limits_{\|x\|=1} \|A\|\ \|Bx\| = \|A\| \max\limits_{\|x\|=1} \|Bx\| = \|A\|\ \|B\|$。

证毕。

称式(1.2)所定义的矩阵范数为从属于所给定向量范数的矩阵范数，又称为矩阵的算子范数。设给定的向量范数为 $\|\cdot\|_p$，则从属于向量范数 $\|\cdot\|_p$ 的矩阵范数仍记为 $\|\cdot\|_p$，即

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p$$

其中 $A \in \mathbf{R}^{n\times n}$，$x \in \mathbf{R}^n$，又称 $\|A\|_p$ 为矩阵 A 的 p-范数。

由定理 1.3 可知，矩阵的 p-范数与向量的 p-范数相容，即对任意的 $A \in \mathbf{R}^{n\times n}$ 和任意的 $x \in \mathbf{R}^n$，有

$$\|Ax\|_p \leqslant \|A\|_p \|x\|_p$$

定理 1.4 设 $A=[a_{ij}]\in\mathbf{R}^{n\times n}$，则

$$\|A\|_1=\max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|$$

$$\|A\|_2=\sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$$

$$\|A\|_\infty=\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|$$

其中 $\lambda_{\max}(A^{\mathrm{T}}A)$表示矩阵 $A^{\mathrm{T}}A$ 的最大特征值。

证 对于 1 -范数，设 $\|x\|_1=\sum_{i=1}^{n}|x_i|=1$。矩阵 A 可表示为

$$A=[\alpha_1,\alpha_2,\cdots,\alpha_n]$$

其中 $\alpha_j=(a_{1j},a_{2j},\cdots,a_{nj})^{\mathrm{T}}$。

设 $\|\alpha_r\|_1=\max_{1\leqslant j\leqslant n}\|\alpha_j\|_1$，则

$$\|Ax\|_1=\|\sum_{j=1}^{n}x_j\alpha_j\|_1\leqslant\sum_{j=1}^{n}|x_j|\;\|\alpha_j\|_1\leqslant$$
$$(\sum_{j=1}^{n}|x_j|)\max_{1\leqslant j\leqslant n}\|\alpha_j\|_1=\max_{1\leqslant j\leqslant n}\|\alpha_j\|_1$$

取向量 $e_r=(0,\cdots,0,1,0,\cdots,0)^{\mathrm{T}}$，它的元素 1 位于第 r 个分量，显然 $\|e_r\|_1=1$，且

$$\|Ae_r\|_1=\|\alpha_r\|_1=\max_{1\leqslant j\leqslant n}\|\alpha_j\|_1$$

于是有

$$\|A\|_1=\max_{\|x\|_1=1}\|Ax\|_1=\max_{1\leqslant j\leqslant n}\|\alpha_j\|_1=\max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|$$

对于 2 -范数，设向量 $x\in\mathbf{R}^n$ 满足 $\|x\|_2=1$。注意到

$$\|Ax\|_2^2=(Ax)^{\mathrm{T}}(Ax)=x^{\mathrm{T}}A^{\mathrm{T}}Ax$$

因 $A^{\mathrm{T}}A$ 是正定或半正定矩阵，故它的全部特征值 $\lambda_i(i=1,2,\cdots,n)$非负，设

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$$

并设相应的标准正交特征向量为 $u_1,u_2,\cdots,u_n$。因而存在实数 $k_1,k_2,\cdots,k_n$，使

$$x=\sum_{i=1}^{n}k_iu_i$$

并且有

$$\|x\|_2^2=x^{\mathrm{T}}x=\sum_{i=1}^{n}k_i^2=1$$

由此可推出

$$\|Ax\|_2^2=x^{\mathrm{T}}A^{\mathrm{T}}Ax=\sum_{i=1}^{n}\lambda_ik_i^2\leqslant\lambda_1$$

取 $\tilde{x}=u_1$，则有 $\|\tilde{x}\|_2=1$，以及

$$\|A\tilde{x}\|_2^2=u_1^{\mathrm{T}}A^{\mathrm{T}}Au_1=\lambda_1$$

所以

$$\|A\|_2=\max_{\|x\|_2=1}\|Ax\|_2=\sqrt{\lambda_1}=\sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}$$

对于 ∞-范数,设向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 满足 $\|\boldsymbol{x}\|_\infty=1$,又设 $\omega=\max\limits_{1\leqslant i\leqslant n}\sum\limits_{j=1}^{n}|a_{ij}|=\sum\limits_{j=1}^{n}|a_{rj}|$,则

$$\|\boldsymbol{Ax}\|_\infty=\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^{n}a_{ij}x_j\right|\leqslant\max_{1\leqslant i\leqslant n}\left(\sum_{j=1}^{n}|a_{ij}||x_j|\right)\leqslant$$

$$\left(\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|\right)\|\boldsymbol{x}\|_\infty=\omega$$

取向量 $\tilde{\boldsymbol{x}}=(\mathrm{sgn}\,a_{r1},\mathrm{sgn}\,a_{r2},\cdots,\mathrm{sgn}\,a_{rn})^{\mathrm{T}}$,其中 sgn 是符号函数。于是有 $\|\tilde{\boldsymbol{x}}\|_\infty=1$ 以及

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|_\infty=\sum_{j=1}^{n}|a_{rj}|=\omega$$

所以

$$\|\boldsymbol{A}\|_\infty=\max_{\|\boldsymbol{x}\|_\infty=1}\|\boldsymbol{Ax}\|_\infty=\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|$$

证毕。

矩阵范数 $\|\cdot\|_1$,$\|\cdot\|_2$ 和 $\|\cdot\|_\infty$ 又分别称为矩阵的列范数、谱范数和行范数,它们都是常用的矩阵范数。

还有一种常用的矩阵范数,就是

$$\|\boldsymbol{A}\|_{\mathrm{F}}=\sqrt{\sum_{i,j=1}^{n}a_{ij}^2}$$

其中 $\boldsymbol{A}=[a_{ij}]\in\mathbf{R}^{n\times n}$。$\|\cdot\|_{\mathrm{F}}$ 称为 Frobenius(佛罗贝尼乌斯)范数,又称为 Euclid 范数,它也可记为 $\|\cdot\|_{\mathrm{E}}$。可以证明,$\|\cdot\|_{\mathrm{F}}$ 与向量范数 $\|\cdot\|_2$ 相容,即

$$\|\boldsymbol{Ax}\|_2\leqslant\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{x}\|_2,\quad \boldsymbol{A}\in\mathbf{R}^{n\times n},\boldsymbol{x}\in\mathbf{R}^n$$

也可以证明,$\|\cdot\|_{\mathrm{F}}$ 满足矩阵范数定义的四个条件。这些证明留给读者完成。需要指出的是,矩阵范数 $\|\cdot\|_{\mathrm{F}}$ 不从属于任何向量范数,即 $\|\cdot\|_{\mathrm{F}}$ 不是算子范数。

单位矩阵 $\boldsymbol{I}$ 的任何一种算子范数都有

$$\|\boldsymbol{I}\|=\max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{Ix}\|=1$$

定理 1.5 设矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的某种范数 $\|\boldsymbol{A}\|<1$,则 $\boldsymbol{I}\pm\boldsymbol{A}$ 为非奇异矩阵,并且当该种范数为算子范数时,还有

$$\|(\boldsymbol{I}\pm\boldsymbol{A})^{-1}\|\leqslant\frac{1}{1-\|\boldsymbol{A}\|}$$

成立。

证 假定 $\boldsymbol{I}\pm\boldsymbol{A}$ 奇异,则齐次线性方程组 $(\boldsymbol{I}\pm\boldsymbol{A})\boldsymbol{x}=\boldsymbol{0}$ 有非零解,即存在向量 $\tilde{\boldsymbol{x}}\neq\boldsymbol{0}$,使

$$\tilde{\boldsymbol{x}}=\mp\boldsymbol{A}\tilde{\boldsymbol{x}}$$

上式两边同取与所用矩阵范数相容的向量范数,得

$$\|\tilde{\boldsymbol{x}}\|=\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|\leqslant\|\boldsymbol{A}\|\|\tilde{\boldsymbol{x}}\|$$

因 $\|\tilde{\boldsymbol{x}}\|>0$,故由上式得 $\|\boldsymbol{A}\|\geqslant1$。这与已知条件相矛盾,因而 $\boldsymbol{I}\pm\boldsymbol{A}$ 必非奇异。

由 $(\boldsymbol{I}-\boldsymbol{A})(\boldsymbol{I}-\boldsymbol{A})^{-1}=\boldsymbol{I}$ 得

$$(\boldsymbol{I}-\boldsymbol{A})^{-1}=\boldsymbol{I}+\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{A})^{-1}$$

上式两边同取所用算子范数,得

$$\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\| \leqslant \|\boldsymbol{I}\| + \|\boldsymbol{A}\| \, \|(\boldsymbol{I}-\boldsymbol{A})^{-1}\|$$

$$(1-\|\boldsymbol{A}\|)\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\| \leqslant \|\boldsymbol{I}\| = 1$$

因$\|\boldsymbol{A}\|<1$,故得

$$\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\| \leqslant \frac{1}{1-\|\boldsymbol{A}\|}$$

同理可证$\|(\boldsymbol{I}+\boldsymbol{A})^{-1}\| \leqslant \frac{1}{1-\|\boldsymbol{A}\|}$。

证毕。

例 6 设 $\boldsymbol{x}=(3,-5,1)^{\mathrm{T}}$

$$\boldsymbol{A} = \begin{bmatrix} 1 & 5 & -2 \\ -2 & 1 & 0 \\ 3 & -8 & 2 \end{bmatrix}$$

试求:$\|\boldsymbol{x}\|_p$, $\|\boldsymbol{A}\|_p(p=1,2,\infty)$以及$\|\boldsymbol{A}\|_{\mathrm{F}}$。

解 $\|\boldsymbol{x}\|_1=3+5+1=9$

$$\|\boldsymbol{x}\|_2 = \sqrt{9+25+1} = \sqrt{35}$$

$$\|\boldsymbol{x}\|_\infty = \max(3,5,1) = 5$$

$$\|\boldsymbol{A}\|_1 = \max(1+2+3,5+1+8,2+0+2) = 14$$

$$\|\boldsymbol{A}\|_\infty = \max(1+5+2,2+1+0,3+8+2) = 13$$

$$\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{1+25+4+4+1+0+9+64+4} = \sqrt{112}$$

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \begin{bmatrix} 14 & -21 & 4 \\ -21 & 90 & -26 \\ 4 & -26 & 8 \end{bmatrix}$$

的特征方程

$$\det(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}-\lambda\boldsymbol{I}) = -\lambda^3 + 112\lambda^2 - 959\lambda + 16 = 0$$

的最大根为$\lambda_1 \approx 102.66$,所以

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_1} \approx 10.132$$

习　　题

1. 下列各近似值均有四位有效数字,

$$a = 0.013\,47, \quad b = -12.341, \quad c = -1.200$$

试指出它们的绝对误差限和相对误差限。

2. 下列各近似值的绝对误差限都是 0.000 5,

$$a = -1.000\,31, \quad b = 0.042, \quad c = -0.000\,32$$

试指出它们有几位有效数字。

3. 已知 a 是积分 $\int_0^1 \mathrm{e}^{-x^2}\,\mathrm{d}x$ 的近似值,并且有四位有效数字,试求 a 的绝对误差限。

4. 已知 $x=2.14\pm0.005$,$y=-1.231\pm0.000\,5$,

(1) 用 $\tilde{u}=\sqrt{2.14}$作为 $u=\sqrt{x}$的近似值;

(2) 用 $\tilde{u}=2.14\times(-1.231)+\frac{2.14}{1.231}$ 作为 $u=xy-\frac{x}{y}$ 的近似值。

试求 $\tilde{u}$ 的绝对误差限和相对误差限，并指出 $\tilde{u}$ 能有几位有效数字。

5. 设 $x=3.214$ 和 $y=3.213$ 都是准确值，在四位十进制限制下，欲计算下列各值

$$u=\sqrt{x}-\sqrt{y}$$

$$u=\tan x-\tan y$$

试选择精确度较高的算法，计算出 u 的近似值。

6. 取 $\sqrt{2}\approx 1.4$，欲计算 $(\sqrt{2}-1)^6$ 的近似值，又已知

$$(\sqrt{2}-1)^6=(3-2\sqrt{2})^3=99-70\sqrt{2}=$$

$$\frac{1}{(\sqrt{2}+1)^6}=\frac{1}{(3+2\sqrt{2})^3}=\frac{1}{99+70\sqrt{2}}$$

试分析这六个公式中使用哪一个进行计算使得误差最小。

7. 在四位十进制的限制下，试选择精确度最高的算法，计算

$$u=1\,340\times 10^2+40+50+60+90$$

的值。

8. 设 $y_n=\int_0^1\frac{x^n}{1+4x}\mathrm{d}x$，在四位十进制的限制下，试使用一个具有数值稳定性的算法，计算 $y_n(n=0,1,\cdots,8)$ 的近似值。

9. 对 $\mathbf{R}^n$ 中任一向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，记 $\|\boldsymbol{x}\|_1=\sum_{i=1}^{n}|x_i|$，$\|\boldsymbol{x}\|_\infty=\max_{1\leqslant i\leqslant n}|x_i|$，试证：$\|\cdot\|_1$ 和 $\|\cdot\|_\infty$ 都是向量范数。

10. 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 是给定的矩阵，对任意的 $\boldsymbol{x}\in\mathbf{R}^n$，记 $\|\boldsymbol{x}\|_{\boldsymbol{A}}=\|\boldsymbol{Ax}\|$，其中 $\|\cdot\|$ 为某种向量范数，试证：若 $\boldsymbol{A}$ 非奇异，则 $\|\cdot\|_{\boldsymbol{A}}$ 是一种向量范数；若 $\boldsymbol{A}$ 奇异，则 $\|\cdot\|_{\boldsymbol{A}}$ 不是向量范数。

11. 设 $\boldsymbol{A}=[a_{ij}]\in\mathbf{R}^{n\times n}$，记

$$\|\boldsymbol{A}\|_{\mathrm{F}}=\sqrt{\sum_{i,j=1}^{n}a_{ij}^2}$$

试证：$\|\cdot\|_{\mathrm{F}}$ 是与向量范数 $\|\cdot\|_2$ 相容的矩阵范数。

12. 已知

$$\boldsymbol{A}=\begin{bmatrix}2&-5&4\\-1&0&3\\4&2&-2\end{bmatrix},\quad \boldsymbol{x}=\begin{bmatrix}4\\-8\\2\end{bmatrix}$$

试求：$\|\boldsymbol{x}\|_p(p=1,2,\infty)$ 以及 $\|\boldsymbol{A}\|_1$，$\|\boldsymbol{A}\|_\infty$，$\|\boldsymbol{A}\|_{\mathrm{F}}$。

13. 证明：对任何非奇异矩阵 $\boldsymbol{A}$，任何矩阵范数，下列不等式成立：

$$\|\boldsymbol{I}\|\geqslant 1,\quad \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|\geqslant 1$$

其中 $\boldsymbol{I}$ 是单位矩阵。

14. 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，$\boldsymbol{x}\in\mathbf{R}^n$，证明：

(1) $\|\boldsymbol{x}\|_2\leqslant\|\boldsymbol{x}\|_1\leqslant\sqrt{n}\|\boldsymbol{x}\|_2$；

(2) $\|\boldsymbol{Ax}\|_2\leqslant\sqrt{n}\|\boldsymbol{A}\|_1\|\boldsymbol{x}\|_2$。

15. 对任意一种矩阵范数，总存在一种与该矩阵范数相容的向量范数，试证明之。

第 2 章　线性方程组的解法

设有 n 元线性方程组

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{2.1}$$

或记为

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{2.2}$$

其中

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

设系数矩阵 $\boldsymbol{A}$ 非奇异，即 $\det \boldsymbol{A} \neq 0$，则方程组(2.1)有唯一解向量 $\boldsymbol{x}$。

求解线性方程组的方法可分为两大类：直接方法和迭代方法（简称迭代法）。本章前几节讨论直接方法，迭代法在最后一节讨论。直接方法的特点是，运用此类方法求解线性方程组时，如果计算过程没有舍入误差，那么经过有限次运算就能求出方程组(2.1)的精确解。

Cramer（克莱姆）法则是直接方法中的一种，根据此法则，方程组(2.1)的解为

$$x_i = \frac{\Delta_i}{\Delta} \quad (i = 1,2,\cdots,n)$$

其中 $\Delta = \det \boldsymbol{A}$，$\Delta_i$ 是 $\boldsymbol{A}$ 中第 i 列换成向量 $\boldsymbol{b}$ 所得矩阵之行列式。假定采用按某行（或某列）展开的方法计算行列式，那么，用 Cramer 法则求解一个 n 元线性方程组所需的乘法运算次数超过 $(n+1)!$。当 n 稍大时，其运算量非常大。Cramer 法则是一个效率低、经济效益差的算法，在实际工作中很少使用。本章将介绍其他常用的直接方法。

2.1　Gauss 消去法

Gauss（高斯）消去法由消元和回代两个过程组成。消元过程就是对方程组(2.1)的增广矩阵

$$[\boldsymbol{A},\boldsymbol{b}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & b_n \end{bmatrix} \tag{2.3}$$

做有限次的初等行变换，使它的系数矩阵部分变换为上三角矩阵。所用的初等行变换主要有两种：第一种，交换两行的位置；第二种，用一个数乘某一行加到另一行上。经过 $k-1$ 次消元后，原增广矩阵(2.3)被变换为

$$[\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)}]=\begin{bmatrix} a_{11}^{(1)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ & \ddots & & & & \vdots & \vdots \\ & & a_{k-1,k-1}^{(k-1)} & \cdots & \cdots & a_{k-1,n}^{(k-1)} & b_{k-1}^{(k-1)} \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & b_k^{(k)} \\ \boldsymbol{O} & & & \vdots & & \vdots & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{bmatrix} \tag{2.4}$$

最后，经过 $n-1$ 次消元，得到

$$[\boldsymbol{A}^{(n)},\boldsymbol{b}^{(n)}]=\begin{bmatrix} a_{11}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ & & \ddots & \vdots & \vdots \\ \boldsymbol{O} & & & a_{nn}^{(n)} & b_n^{(n)} \end{bmatrix}$$

其中 $a_{ii}^{(i)}(i=1,2,\cdots,n)$不为零。以$[\boldsymbol{A}^{(n)},\boldsymbol{b}^{(n)}]$作为增广矩阵的上三角线性方程组

$$\begin{cases} a_{11}^{(1)}x_1+a_{12}^{(1)}x_2+\cdots+a_{1n}^{(1)}x_n=b_1^{(1)} \\ \qquad a_{22}^{(2)}x_2+\cdots+a_{2n}^{(2)}x_n=b_2^{(2)} \\ \qquad\qquad \vdots \\ \qquad\qquad\qquad a_{nn}^{(n)}x_n=b_n^{(n)} \end{cases} \tag{2.5}$$

与原方程组(2.1)是同解方程组。回代过程就是先由方程组(2.5)的最后一个方程解出 x_n，然后通过逐步回代，依次求出 $x_{n-1},x_{n-2},\cdots,x_1$。这种 Gauss 消去法可分为顺序 Gauss 消去法和列主元素 Gauss 消去法两种。

2.1.1　顺序 Gauss 消去法

在 Gauss 消去法的消元过程中对方程组的增广矩阵只做前述的第二种初等行变换就形成了顺序 Gauss 消去法，其算法如下：

记

$$a_{ij}^{(1)}=a_{ij}\quad(i,j=1,2,\cdots,n)$$
$$b_i^{(1)}=b_i\quad(i=1,2,\cdots,n)$$

1. 消元过程

对于 $k=1,2,\cdots,n-1$ 执行

(1) 如果 $a_{kk}^{(k)}=0$，则算法失效，停止计算；否则转(2)。

(2) 对于 $i=k+1,k+2,\cdots,n$ 计算

$$m_{ik}=a_{ik}^{(k)}/a_{kk}^{(k)}$$
$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}\quad(j=k+1,k+2,\cdots,n)$$
$$b_i^{(k+1)}=b_i^{(k)}-m_{ik}b_k^{(k)}$$

2. 回代过程

$$x_n=b_n^{(n)}/a_{nn}^{(n)}$$
$$x_k=\left(b_k^{(k)}-\sum_{j=k+1}^{n}a_{kj}^{(k)}x_j\right)/a_{kk}^{(k)}\quad(k=n-1,n-2,\cdots,1)$$

由上述算法可统计出顺序 Gauss 消去法求解 n 元线性方程组的乘除法运算总次数为 $\frac{1}{3}(n^3+3n^2-n)$。与 Cramer 法则相比，顺序 Gauss 消去法的计算量大为减少。例如，当 $n=20$ 时，Cramer 法则的乘除法运算次数超过 5×10^{19} 次，而顺序 Gauss 消去法只需3 060次。

顺序 Gauss 消去法计算过程中出现的 $a_{kk}^{(k)}(k=1,2,\cdots,n)$ 称为主元素。它们是由原始增广矩阵$[\boldsymbol{A},\boldsymbol{b}]$按自然顺序消元时产生的。即使 $\det \boldsymbol{A}\neq0$，也可能对某个 $k<n$，出现 $a_{kk}^{(k)}=0$。这时，消元过程就进行不下去了。

定理 2.1 顺序 Gauss 消去法的前 $n-1$ 个主元素 $a_{kk}^{(k)}(k=1,2,\cdots,n-1)$ 均不为零的充分必要条件是方程组(2.1)的系数矩阵 $\boldsymbol{A}$ 的前 $n-1$ 个顺序主子式

$$D_k=\begin{vmatrix} a_{11}^{(1)} & \cdots & a_{1k}^{(1)} \\ \vdots & & \vdots \\ a_{k1}^{(1)} & \cdots & a_{kk}^{(1)} \end{vmatrix}\neq 0 \quad (k=1,2,\cdots,n-1) \tag{2.6}$$

证

充分性　设条件(2.6)成立。因 $D_1=a_{11}^{(1)}$，故 $a_{11}^{(1)}\neq0$，因而可作顺序 Gauss 消去法的第一次消元，产生主元素 $a_{22}^{(2)}$。由行列式性质可知

$$D_2=\begin{vmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ 0 & a_{22}^{(2)} \end{vmatrix}=a_{11}^{(1)}a_{22}^{(2)}$$

故 $a_{22}^{(2)}\neq0$。设已经过 $r-2(r\geqslant3)$ 次消元，所产生的主元素 $a_{22}^{(2)},\cdots,a_{r-1,r-1}^{(r-1)}$ 均不为零，则可作顺序 Gauss 消去法的第 $r-1$ 次消元，产生主元素 $a_{rr}^{(r)}$。由行列式性质可知

$$D_r=\begin{vmatrix} a_{11}^{(1)} & \cdots & a_{1r}^{(1)} \\ & \ddots & \vdots \\ & & a_{rr}^{(r)} \end{vmatrix}=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{rr}^{(r)}$$

故 $a_{rr}^{(r)}\neq0$。当 $r=n-1$ 时，就得出 $a_{kk}^{(k)}\neq0(k=1,2,\cdots,n-1)$。

必要性　设 $a_{kk}^{(k)}\neq0(k=1,2,\cdots,n-1)$，则由

$$D_k=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{kk}^{(k)} \quad (k=1,2,\cdots,n-1)$$

可知 $D_k\neq0(k=1,2,\cdots,n-1)$。

证毕。

当方程组(2.1)的系数矩阵 $\boldsymbol{A}$ 非奇异，并且 $\boldsymbol{A}$ 的前 $n-1$ 个顺序主子式均不为零，就可以使用顺序 Gauss 消去法求解。但是，如果在求解过程中遇到某个 k，虽然 $a_{kk}^{(k)}\neq0$ 但 $|a_{kk}^{(k)}|$ 很小，使 $-m_{ik}$(称为行乘数)的绝对值很大，那么，舍入误差的积累就会很大。计算出的解相对于精确解会有很大的误差。因此，顺序 Gauss 消去法的数值稳定性是没有保证的。

2.1.2 列主元素 Gauss 消去法

在 Gauss 消去法的消元过程中，第 k 次消元之前，先对增广矩阵$[\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)}]$[参见式(2.4)]作前述的第一种初等行变换(行交换)，目的是把 $a_{ik}^{(k)}(i=k,k+1,\cdots,n)$ 中绝对值最大的元素交换到第 k 行的主对角线位置上，然后再使用前述的第二种初等行变换进行消元，这就形成了列主元素 Gauss 消去法，其算法如下：

记

$$a_{ij}^{(1)}=a_{ij} \quad (i,j=1,2,\cdots,n)$$

$$b_i^{(1)} = b_i \quad (i = 1,2,\cdots,n)$$

1. 消元过程

对于 $k=1,2,\cdots,n-1$ 执行

(1) 选行号 i_k，使 $|a_{i_k k}^{(k)}| = \max\limits_{k\leqslant i\leqslant n}|a_{ik}^{(k)}|$。

(2) 交换 $a_{kj}^{(k)}$ 与 $a_{i_k j}^{(k)}$ $(j=k,k+1,\cdots,n)$ 以及 $b_k^{(k)}$ 与 $b_{i_k}^{(k)}$ 所含的数值。

(3) 对于 $i=k+1,k+2,\cdots,n$ 计算

$$m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} \quad (j = k+1,k+2,\cdots,n)$$

$$b_i^{(k+1)} = b_i^{(k)} - m_{ik}b_k^{(k)}$$

2. 回代过程

$$x_n = b_n^{(n)}/a_{nn}^{(n)}$$

$$x_k = \left(b_k^{(k)} - \sum_{j=k+1}^{n} a_{kj}^{(k)}x_j\right)/a_{kk}^{(k)} \quad (k = n-1,n-2,\cdots,1)$$

在此算法中的 $a_{i_k k}^{(k)}$ $(k=1,2,\cdots,n-1)$ 称为第 k 个列主元素，它的数值总要被交换到第 k 个主对角线元素的位置上。

定理 2.2　设方程组(2.1)的系数矩阵 $\boldsymbol{A}$ 非奇异，则用列主元素 Gauss 消去法求解方程组(2.1)时，各个列主元素 $a_{i_k k}^{(k)}$ $(k=1,2,\cdots,n-1)$ 均不为零。

证　用反证法。假定存在某个 $r(1\leqslant r\leqslant n-1)$，前 $r-1$ 个列主元素不为零，而 $a_{i_r r}^{(r)}=0$，则有 $a_{ir}^{(r)}=0(i=r,r+1,\cdots,n)$。由行列式性质可知

$$\det\boldsymbol{A} = \pm\begin{vmatrix} a_{11}^{(1)} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ & \ddots & & & & & \vdots \\ & & a_{r-1,r-1}^{(r-1)} & \cdots & \cdots & \cdots & a_{r-1,n}^{(r-1)} \\ & & & 0 & a_{r,r+1}^{(r)} & \cdots & a_{rn}^{(r)} \\ & & & \vdots & \vdots & & \vdots \\ & & & 0 & a_{n,r+1}^{(r)} & \cdots & a_{nn}^{(r)} \end{vmatrix} =$$

$$\pm a_{11}^{(1)}\cdots a_{r-1,r-1}^{(r-1)}\begin{vmatrix} 0 & a_{r,r+1}^{(r)} & \cdots & a_{rn}^{(r)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n,r+1}^{(r)} & \cdots & a_{nn}^{(r)} \end{vmatrix} = 0$$

与 $\boldsymbol{A}$ 非奇异相矛盾，故 $a_{i_k k}^{(k)}$ $(k=1,2,\cdots,n-1)$ 均不为零。

证毕。

用列主元素 Gauss 消去法求解线性方程组(2.1)，不仅只需系数矩阵 $\boldsymbol{A}$ 非奇异，而且对一般线性方程组，此方法还具有良好的数值稳定性，其计算量与顺序 Gauss 消去法相同。

例 1　在四位十进制的限制下，试分别用顺序 Gauss 消去法和列主元素 Gauss 消去法求解下列线性方程组

$$\begin{cases} 0.012\,x_1 + 0.01\,x_2 + 0.167\,x_3 = 0.678\,1 \\ x_1 + 0.833\,4\,x_2 + 5.91\,x_3 = 12.1 \\ 3\,200\,x_1 + 1\,200\,x_2 + 4.2\,x_3 = 981 \end{cases}$$

解 用顺序 Gauss 消去法求解，消元过程如下：

$$\begin{bmatrix} 0.01200 & 0.01000 & 0.1670 & 0.6781 \\ 1.000 & 0.8334 & 5.910 & 12.10 \\ 3200 & 1200 & 4.200 & 981.0 \end{bmatrix} \to$$

$$\begin{bmatrix} 0.01200 & 0.01000 & 0.1670 & 0.6781 \\ 0 & 0.1000\times 10^{-3} & -8.010 & -44.41 \\ 0 & -1467 & -4454\times 10 & -1798\times 10^{2} \end{bmatrix} \to$$

$$\begin{bmatrix} 0.01200 & 0.01000 & 0.1670 & 0.6781 \\ 0 & 0.1000\times 10^{-3} & -8.010 & -44.41 \\ 0 & 0 & -1175\times 10^{5} & -6517\times 10^{5} \end{bmatrix}$$

经回代，得

$$x_3 = 5.546,\quad x_2 = 100.0,\quad x_1 = -104.0$$

用列主元素 Gauss 消去法，消元过程如下(带框者为主元素)：

$$\begin{bmatrix} 0.01200 & 0.01000 & 0.1670 & 0.6781 \\ 1.000 & 0.8334 & 5.910 & 12.10 \\ \boxed{3200} & 1200 & 4.200 & 981.0 \end{bmatrix} \to$$

$$\begin{bmatrix} 3200 & 1200 & 4.200 & 981.0 \\ 0 & \boxed{0.4584} & 5.909 & 11.79 \\ 0 & 0.5500\times 10^{-2} & 0.1670 & 0.6744 \end{bmatrix} \to$$

$$\begin{bmatrix} 3200 & 1200 & 4.200 & 981.0 \\ 0 & 0.4584 & 5.909 & 11.79 \\ 0 & 0 & 0.09609 & 0.5329 \end{bmatrix}$$

经回代，得

$$x_3 = 5.546,\quad x_2 = -45.77,\quad x_1 = 17.46$$

本例题的线性方程组的精确解舍入到四位有效数字是

$$x_3 = 5.546,\quad x_2 = -45.76,\quad x_1 = 17.46$$

由此看出，列主元素 Gauss 消去法的精度显著高于顺序 Gauss 消去法。对于此例，由于顺序 Gauss 消去法中的主元素绝对值非常小，使行乘数绝对值非常大，出现较小数加不到较大数中的现象，舍入误差的积累很大，所得结果中 $x_1=-104.0$ 和 $x_2=100.0$ 完全失真。

2.2 直接三角分解法

2.2.1 Doolittle 分解法与 Crout 分解法

如果方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 能分解成

$$\boldsymbol{A} = \boldsymbol{LU} \tag{2.7}$$

其中 $\boldsymbol{L}$ 是下三角矩阵，$\boldsymbol{U}$ 是上三角矩阵。这时，方程组(2.2)就可化为两个容易求解的三角形方程组

$$\boldsymbol{Ly} = \boldsymbol{b},\quad \boldsymbol{Ux} = \boldsymbol{y}$$

先由 $\boldsymbol{Ly}=\boldsymbol{b}$ 解出向量 $\boldsymbol{y}$，再由 $\boldsymbol{Ux}=\boldsymbol{y}$ 解出向量 $\boldsymbol{x}$，这就是原方程组(2.2)的解向量。

矩阵 $\boldsymbol{A}$ 分解为式(2.7)的形式称为矩阵 $\boldsymbol{A}$ 的三角分解。如果在分解式(2.7)中 $\boldsymbol{L}$ 是单位下三角矩阵，$\boldsymbol{U}$ 是上三角矩阵，则式(2.7)又称为矩阵 $\boldsymbol{A}$ 的 Doolittle(杜利特尔)分解；如果 $\boldsymbol{L}$ 是下三角矩阵，$\boldsymbol{U}$ 是单位上三角矩阵，则式(2.7)又称为矩阵 $\boldsymbol{A}$ 的 Crout(克劳特)分解。矩阵能作三角分解是有条件的。

定义　称 $n\times n$ 矩阵

$$\boldsymbol{P}_k=\begin{bmatrix}1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & p_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & p_{n,k} & & & 1\end{bmatrix}\quad (k=1,2,\cdots,n-1) \tag{2.8}$$

为初等下三角矩阵。

定理 2.3　矩阵 $\boldsymbol{A}=[a_{ij}]_{n\times n}(n\geqslant 2)$ 有唯一的 Doolittle 分解的充分必要条件是 $\boldsymbol{A}$ 的前 $n-1$ 个顺序主子式 $D_k\neq 0(k=1,2,\cdots,n-1)$。

证

充分性　因 $D_k\neq 0\ (k=1,2,\cdots,n-1)$，根据定理 2.1，可对矩阵 $\boldsymbol{A}$ 进行顺序 Gauss 消去法中的初等行变换，把 $\boldsymbol{A}$ 变换为上三角矩阵，其变换过程相当于

$$\boldsymbol{P}_{n-1}\cdots\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{A}=\boldsymbol{U} \tag{2.9}$$

其中 $\boldsymbol{P}_k(k=1,2,\cdots,n-1)$ 为式(2.8)的初等下三角矩阵，$\boldsymbol{P}_k$ 中的 $p_{ik}(i=k+1,\cdots,n)$ 是顺序 Gauss 消去法中的行乘数 $-m_{ik}$；$\boldsymbol{U}$ 为上三角矩阵，并且其主对角线元素 $u_{kk}=a_{kk}^{(k)}\neq 0\ (k=1,2,\cdots,n-1)$。若 $\boldsymbol{A}$ 非奇异，则 $u_{nn}\neq 0$；若 $\boldsymbol{A}$ 奇异，则 $u_{nn}=0$。

由式(2.9)得

$$\boldsymbol{A}=\boldsymbol{P}_1^{-1}\boldsymbol{P}_2^{-1}\cdots\boldsymbol{P}_{n-1}^{-1}\boldsymbol{U}$$

记 $\boldsymbol{L}=\boldsymbol{P}_1^{-1}\boldsymbol{P}_2^{-1}\cdots\boldsymbol{P}_{n-1}^{-1}$，$\boldsymbol{L}$ 是单位下三角矩阵，于是有

$$\boldsymbol{A}=\boldsymbol{LU}$$

现在证明唯一性。设 $\boldsymbol{A}$ 有两种 Doolittle 分解

$$\boldsymbol{A}=\boldsymbol{LU}=\boldsymbol{L}^*\boldsymbol{U}^* \tag{2.10}$$

当 $\boldsymbol{A}$ 非奇异时，$\boldsymbol{U}$ 和 $\boldsymbol{U}^*$ 都非奇异，由式(2.10)得

$$\boldsymbol{UU}^{*-1}=\boldsymbol{L}^{-1}\boldsymbol{L}^*$$

由于 $\boldsymbol{L}^{-1}\boldsymbol{L}^*$ 是单位下三角矩阵，$\boldsymbol{UU}^{*-1}$ 是上三角矩阵，所以，它们只能是单位矩阵，即

$$\boldsymbol{UU}^{*-1}=\boldsymbol{L}^{-1}\boldsymbol{L}^*=\boldsymbol{I}$$

因而 $\boldsymbol{U}=\boldsymbol{U}^*$，$\boldsymbol{L}=\boldsymbol{L}^*$。

当 $\boldsymbol{A}$ 奇异时，$\boldsymbol{U}$ 和 $\boldsymbol{U}^*$ 都奇异，且它们的主对角线元素 u_{ii} 和 u_{ii}^* 满足 $u_{ii}\neq 0, u_{ii}^*\neq 0(i=1,2,\cdots,n-1)$，$u_{nn}=0, u_{nn}^*=0$。把 $\boldsymbol{LU}=\boldsymbol{L}^*\boldsymbol{U}^*$ 写成

$$\begin{bmatrix}\boldsymbol{L}_{n-1} & \boldsymbol{O}\\ \boldsymbol{r}^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_{n-1} & \boldsymbol{s}\\ \boldsymbol{O} & 0\end{bmatrix}=\begin{bmatrix}\boldsymbol{L}_{n-1}^* & \boldsymbol{O}\\ \boldsymbol{r}^{*\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_{n-1}^* & \boldsymbol{s}^*\\ \boldsymbol{O} & 0\end{bmatrix}$$

由此可知

$$\boldsymbol{L}_{n-1}\boldsymbol{U}_{n-1}=\boldsymbol{L}_{n-1}^*\boldsymbol{U}_{n-1}^*,\quad \boldsymbol{r}^{\mathrm{T}}\boldsymbol{U}_{n-1}=\boldsymbol{r}^{*\mathrm{T}}\boldsymbol{U}_{n-1}^*,\quad \boldsymbol{L}_{n-1}\boldsymbol{s}=\boldsymbol{L}_{n-1}^*\boldsymbol{s}^*$$

由于$\boldsymbol{U}_{n-1}$和$\boldsymbol{U}_{n-1}^*$非奇异，故有

$$\boldsymbol{L}_{n-1}=\boldsymbol{L}_{n-1}^*,\quad \boldsymbol{U}_{n-1}=\boldsymbol{U}_{n-1}^*,\quad \boldsymbol{r}^{\mathrm{T}}=\boldsymbol{r}^{*\mathrm{T}},\quad \boldsymbol{s}=\boldsymbol{s}^*$$

因而$\boldsymbol{U}=\boldsymbol{U}^*$，$\boldsymbol{L}=\boldsymbol{L}^*$。

必要性 设矩阵$\boldsymbol{A}$有唯一的Doolittle分解$\boldsymbol{A}=\boldsymbol{LU}$，此时必有$u_{ii}\neq 0\ (i=1,2,\cdots,n-1)$；否则就存在$u_{kk}=0(1\leqslant k\leqslant n-1)$，而$u_{11},u_{22},\cdots,u_{k-1,k-1}$不为零。那么由

$$\boldsymbol{A}_{k+1}=\begin{bmatrix}\boldsymbol{A}_k & \boldsymbol{y}\\ \boldsymbol{x}^{\mathrm{T}} & a_{k+1,k+1}\end{bmatrix}=\begin{bmatrix}\boldsymbol{L}_k & \boldsymbol{O}\\ \boldsymbol{r}^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_k & \boldsymbol{s}\\ \boldsymbol{O} & u_{k+1,k+1}\end{bmatrix}$$

（其中$\boldsymbol{A}_k$，$\boldsymbol{L}_k$和$\boldsymbol{U}_k$分别是$\boldsymbol{A}$，$\boldsymbol{L}$和$\boldsymbol{U}$的k阶顺序主子矩阵）可知

$$\boldsymbol{x}^{\mathrm{T}}=\boldsymbol{r}^{\mathrm{T}}\boldsymbol{U}_k,\quad \boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{r}=\boldsymbol{x}$$

因$\boldsymbol{U}_k$奇异，故$\boldsymbol{r}$不存在或存在不唯一。这与矩阵$\boldsymbol{A}$有唯一的Doolittle分解相矛盾。

由$u_{ii}\neq 0(i=1,2,\cdots,n-1)$以及$\boldsymbol{A}_k=\boldsymbol{L}_k\boldsymbol{U}_k$可知

$$D_k=\det \boldsymbol{A}_k=u_{11}u_{22}\cdots u_{kk}\neq 0\quad (k=1,2,\cdots,n-1)$$

证毕。

推论 矩阵$\boldsymbol{A}=[a_{ij}]_{n\times n}(n\geqslant 2)$有唯一的Crout分解的充分必要条件是$\boldsymbol{A}$的前$n-1$个顺序主子式$D_k\neq 0\ (k=1,2,\cdots,n-1)$。

证 只须证明，$\boldsymbol{A}$有唯一的Doolittle分解与$\boldsymbol{A}$有唯一的Crout分解是等价的。事实上，当$\boldsymbol{A}$有唯一的Doolittle分解$\boldsymbol{A}=\boldsymbol{LU}$时，$\boldsymbol{U}$的前$n-1$个主对角线元素$u_{ii}\neq 0(i=1,2,\cdots,n-1)$，因而有

$$\boldsymbol{U}=\boldsymbol{D}\tilde{\boldsymbol{U}}$$

其中$\boldsymbol{D}=\mathrm{diag}(u_{11},u_{22},\cdots,u_{nn})$，$\tilde{\boldsymbol{U}}$是单位上三角矩阵，且$\boldsymbol{D}$和$\tilde{\boldsymbol{U}}$都是唯一的。于是有

$$\boldsymbol{A}=\boldsymbol{LU}=\boldsymbol{LD}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$$

其中$\tilde{\boldsymbol{L}}=\boldsymbol{LD}$是下三角矩阵，$\tilde{\boldsymbol{L}}$也是唯一的。反之，当$\boldsymbol{A}$有唯一的Crout分解$\boldsymbol{A}=\tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$时，可以证明，$\tilde{\boldsymbol{L}}$的前$n-1$个主对角线元素$\tilde{l}_{ii}\neq 0\ (i=1,2,\cdots,n-1)$。由此可推出，$\boldsymbol{A}$必有唯一的Doolittle分解$\boldsymbol{A}=\boldsymbol{LU}$。

证毕。

设线性方程组(2.2)的系数矩阵$\boldsymbol{A}=[a_{ij}]_{n\times n}$非奇异，且$\boldsymbol{A}$的前$n-1$个顺序主子式都不为零。对$\boldsymbol{A}$作Doolittle分解

$$\boldsymbol{A}=\boldsymbol{LU}=\begin{bmatrix}1 & & & \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1\end{bmatrix}\begin{bmatrix}u_{11} & u_{12} & \cdots & u_{1n}\\ & u_{22} & \cdots & u_{2n}\\ & & \ddots & \vdots\\ & & & u_{nn}\end{bmatrix}\tag{2.11}$$

由分解式(2.11)以及矩阵相乘法，可知

$$a_{1j}=u_{1j}\qquad (j=1,2,\cdots,n)$$
$$a_{i1}=l_{i1}u_{11}\qquad (i=2,3,\cdots,n)$$

当$k=2,3,\cdots,n$时，有

$$a_{kj}=\sum_{t=1}^{n}l_{kt}u_{tj}=\sum_{t=1}^{k-1}l_{kt}u_{tj}+u_{kj}\quad (j=k,k+1,\cdots,n)$$

$$a_{ik}=\sum_{t=1}^{n}l_{it}u_{tk}=\sum_{t=1}^{k-1}l_{it}u_{tk}+l_{ik}u_{kk}\quad (i=k+1,k+2,\cdots,n;\ k<n)$$

由上述各式得到矩阵 $A=[a_{ij}]_{n\times n}$ 的 Doolittle 分解计算公式：

$$\begin{cases}\text{对于 } k=1,2,\cdots,n \text{ 计算} \\ u_{kj}=a_{kj}-\sum_{t=1}^{k-1} l_{kt}u_{tj} \quad (j=k,k+1,\cdots,n) \\ l_{ik}=(a_{ik}-\sum_{t=1}^{k-1} l_{it}u_{tk})/u_{kk} \quad (i=k+1,k+2,\cdots,n;\ k<n)\end{cases} \tag{2.12}$$

在计算机上计算时，可把 u_{kj} 和 l_{ik} 的数值分别存放在原 a_{kj} 和 a_{ik} 的存储单元内。分解计算完后，原矩阵 A 就成为

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ l_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ l_{n1} & \cdots & l_{n,n-1} & u_{nn}\end{bmatrix}$$

求解下三角方程组 $Ly=b$ 和上三角方程组 $Ux=y$ 的计算公式为

$$\begin{cases} y_1=b_1 \\ y_i=b_i-\sum_{t=1}^{i-1} l_{it}y_t \quad (i=2,3,\cdots,n) \\ x_n=y_n/u_{nn} \\ x_i=(y_i-\sum_{t=i+1}^{n} u_{it}x_t)/u_{ii} \quad (i=n-1,n-2,\cdots,1)\end{cases} \tag{2.13}$$

计算公式(2.12)、(2.13)就是求解方程组(2.2)的 Doolittle 分解法。

同理，可推出矩阵 $A=[a_{ij}]_{n\times n}$ 的 Crout 分解

$$A=\tilde{L}\tilde{U}=\begin{bmatrix} l_{11} & & \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn}\end{bmatrix}\begin{bmatrix} 1 & u_{12} & \cdots & u_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & u_{n-1,n} \\ & & & 1\end{bmatrix}$$

的计算公式：

$$\begin{cases}\text{对于 } k=1,2,\cdots,n \text{ 计算} \\ l_{ik}=a_{ik}-\sum_{t=1}^{k-1} l_{it}u_{tk} \quad (i=k,k+1,\cdots,n) \\ u_{kj}=(a_{kj}-\sum_{t=1}^{k-1} l_{kt}u_{tj})/l_{kk} \quad (j=k+1,k+2,\cdots,n;\ k<n)\end{cases} \tag{2.14}$$

以及求解下三角方程组 $\tilde{L}y=b$ 和上三角方程组 $\tilde{U}x=y$ 的计算公式：

$$\begin{cases} y_1=b_1/l_{11} \\ y_i=(b_i-\sum_{t=1}^{i-1} l_{it}y_t)/l_{ii} \quad (i=2,3,\cdots,n) \\ x_n=y_n \\ x_i=y_i-\sum_{t=i+1}^{n} u_{it}x_t \quad (i=n-1,n-2,\cdots,1)\end{cases} \tag{2.15}$$

计算公式(2.14)、(2.15)就是求解方程组(2.2)的 Crout 分解法。

例 2 在四位十进制的限制下，用 Doolittle 分解法求解下列方程组

$$\begin{cases} 8.1x_1 + 2.3x_2 - 1.5x_3 = 6.1 \\ 0.5x_1 - 6.23x_2 + 0.87x_3 = 2.3 \\ 2.5x_1 + 1.5x_2 + 10.2x_3 = 1.8 \end{cases}$$

解 此方程组的系数矩阵 $\boldsymbol{A}$ 的各阶顺序主子式不为零，必存在 Doolittle 分解 $\boldsymbol{A}=\boldsymbol{LU}$，且 $\boldsymbol{U}$ 非奇异。利用计算公式(2.12)，分解结果为

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21} & u_{22} & u_{23} \\ l_{31} & l_{32} & u_{33} \end{bmatrix} = \begin{bmatrix} 8.100 & 2.300 & -1.500 \\ 0.06173 & -6.372 & 0.9626 \\ 0.3086 & -0.1240 & 10.78 \end{bmatrix}$$

再利用计算公式(2.13)，得

$$y_1=6.1,\quad y_2=1.923,\quad y_3=0.1560$$

$$x_3=0.01447,\quad x_2=-0.2996,\quad x_1=0.8408$$

此方程组的精确解舍入到四位有效数字是

$$x_3=0.01445,\quad x_2=-0.2997,\quad x_1=0.8409$$

2.2.2 选主元的 Doolittle 分解法

定义 称 $n\times n$ 矩阵

$$\boldsymbol{Q}_k = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & & & 1 & & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & 1 & & & 0 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \text{第 } k \text{ 行} \\ \\ \\ \\ \text{第 } i_k \text{ 行} \\ \\ \\ \\ \end{matrix} \tag{2.16}$$

为初等置换矩阵，称每一行和每一列都只有一个非零元素 1 的$n\times n$矩阵为置换矩阵。

设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$，则 $\boldsymbol{Q}_k\boldsymbol{A}$ 相当于交换 $\boldsymbol{A}$ 的第 k 行和第 i_k 行的位置，$\boldsymbol{A}\boldsymbol{Q}_k$ 相当于交换 $\boldsymbol{A}$ 的第 k 列和第 i_k 列的位置。若干个初等置换矩阵的乘积 $\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_r=\boldsymbol{Q}$ 是置换矩阵。

定理 2.4 若矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$非奇异，则存在置换矩阵 $\boldsymbol{Q}$，使 $\boldsymbol{QA}$ 可作 Doolittle 分解

$$\boldsymbol{QA} = \boldsymbol{LU}$$

其中 $\boldsymbol{L}$ 是单位下三角矩阵，$\boldsymbol{U}$ 是上三角矩阵。

证 因 $\boldsymbol{A}$ 非奇异，根据定理 2.2，可对 $\boldsymbol{A}$ 进行列主元素 Gauss 消去法中的初等行变换，把 $\boldsymbol{A}$ 变换为上三角矩阵 $\boldsymbol{U}$。其变换过程相当于

$$\boldsymbol{P}_{n-1}\boldsymbol{Q}_{n-1}\cdots\boldsymbol{P}_2\boldsymbol{Q}_2\boldsymbol{P}_1\boldsymbol{Q}_1\boldsymbol{A} = \boldsymbol{U} \tag{2.17}$$

其中 $\boldsymbol{P}_k$ 和 $\boldsymbol{Q}_k(k=1,2,\cdots,n-1)$分别是式(2.8)和式(2.16)所示的初等下三角矩阵[其中$p_{ik}=-m_{ik}(i=k+1,\cdots,n)$]和初等置换矩阵。

下面以 $n=4$ 为例继续证明。因 $Q_kQ_k=I$(单位矩阵)，故 $n=4$ 的式(2.17)可写成

$$P_3(Q_3P_2Q_3)(Q_3Q_2P_1Q_2Q_3)Q_3Q_2Q_1A=U$$

$$P_3\widetilde{P}_2\widetilde{P}_1Q_3Q_2Q_1A=U$$

其中 $\widetilde{P}_1=Q_3Q_2P_1Q_2Q_3$ 和 $\widetilde{P}_2=Q_3P_2Q_3$ 都是形如式(2.8)的初等下三角矩阵。于是有

$$QA=LU$$

其中 $Q=Q_3Q_2Q_1$ 是置换矩阵，$L=\widetilde{P}_1^{-1}\widetilde{P}_2^{-1}P_3^{-1}$ 是单位下三角矩阵。

证毕。

定理 2.4 说明，只要矩阵 A 非奇异，则通过对 A 作适当的行变换就可进行 Doolittle 分解，而不必要求 A 的前 $n-1$ 个顺序主子式都不为零。

设方程组(2.2)的系数矩阵 A 非奇异。当用 Doolittle 分解法(2.12)、(2.13)求解方程组(2.2)时，有可能出现某个 $k(1\leqslant k\leqslant n-1)$，使 $u_{kk}=0$，或者虽然 $u_{kk}\neq 0$，但 $|u_{kk}|$ 相对很小。如果是前者，则 A 的分解计算不能按公式(2.12)的算法进行下去；如果是后者，虽然能按该算法继续分解计算，但会引起很大的舍入误差积累。因此，为了提高求解的精度，可采用与列主元素 Gauss 消去法类似的方法，在 Doolittle 分解中也选主元素。根据定理 2.4，可通过对矩阵 A 进行行变换，实现分解 $QA=LU$，并通过行变换实现选主元素，使得不仅 $u_{kk}\neq 0(k=1,2,\cdots,n-1)$，而且 $|u_{kk}|$ 尽量大一些。设按公式(2.12)的分解已进行了 $k-1$ 步，原来存放 $a_{ij}(i,j=1,2,\cdots,n)$ 的矩阵已成为

$$A^{(k-1)}=\begin{bmatrix} u_{11} & \cdots & \cdots & \cdots & u_{1k} & \cdots & u_{1n} \\ l_{21} & u_{22} & \cdots & \cdots & u_{2k} & \cdots & u_{2n} \\ \vdots & & \ddots & & \vdots & & \vdots \\ & & & u_{k-1,k-1} & u_{k-1,k} & \cdots & u_{k-1,n} \\ l_{k1} & \cdots & \cdots & l_{k,k-1} & a_{kk} & \cdots & a_{kn} \\ \vdots & & & \vdots & \vdots & & \vdots \\ l_{n1} & \cdots & \cdots & l_{n,k-1} & a_{nk} & \cdots & a_{nn} \end{bmatrix}$$

在第 k 步，先计算中间量

$$s_i=a_{ik}-\sum_{t=1}^{k-1}l_{it}u_{tk}\quad(i=k,k+1,\cdots,n)$$

满足 $|s_{i_k}|=\max\limits_{k\leqslant i\leqslant n}|s_i|$ 的 s_{i_k} 就是第 k 步的主元素，应以主元素 s_{i_k} 的值作为 u_{kk}。为此，只须交换矩阵 $A^{(k-1)}$ 的第 k 行与第 i_k 行元素所含的数值，再按公式(2.12)的算法进行第 k 步的分解计算。这相当于先交换原矩阵 A 的第 k 行与第 i_k 行，得到 QA，再对 QA 进行 Doolittle 分解 $QA=LU$。这时，原方程组(2.2)成为 $LUx=Qb$，可化为求解两个三角方程组

$$Ly=Qb,\quad Ux=y$$

上述求解方程组(2.2)的方法称为选主元的 Doolittle 分解法，其具体算法如下：

(1) 作分解 $QA=LU$。

设置整型数组 $M(n)$，它的第 k 个元素 M_k 用于纪录第 k 个主元素所在的行号。

对于 $k=1,2,\cdots,n$ 执行

① 计算中间量

$$s_i=a_{ik}-\sum_{t=1}^{k-1}l_{it}u_{tk}\quad(i=k,k+1,\cdots,n)$$

② 选行号 i_k，使 $|s_{i_k}|=\max\limits_{k\leqslant i\leqslant n}|s_i|$，令 $M_k=i_k$。

③ 若 $i_k=k$，则转④；否则交换 l_{kt} 与 $l_{i_k t}(t=1,2,\cdots,k-1)$、$a_{kt}$ 与 $a_{i_k t}(t=k,k+1,\cdots,n)$ 以及 s_k 与 s_{i_k} 所含的数值，转④。

④ 计算

$$u_{kk}=s_k$$

$$u_{kj}=a_{kj}-\sum_{t=1}^{k-1}l_{kt}u_{tj}\quad(j=k+1,k+2,\cdots,n;\ k<n)$$

$$l_{ik}=s_i/u_{kk}\quad(i=k+1,k+2,\cdots,n;\ k<n)$$

(2) 求 $\boldsymbol{Qb}$。

对于 $k=1,2,\cdots,n-1$ 执行

① $t=M_k$。

② 交换 b_k 与 b_t 所含的数值。

(3) 求解 $\boldsymbol{Ly}=\boldsymbol{Qb}$ 和 $\boldsymbol{Ux}=\boldsymbol{y}$

$$y_1=b_1$$

$$y_i=b_i-\sum_{t=1}^{i-1}l_{it}y_t\quad(i=2,3,\cdots,n)$$

$$x_n=y_n/u_{nn}$$

$$x_i=\Big(y_i-\sum_{t=i+1}^{n}u_{it}x_t\Big)/u_{ii}\quad(i=n-1,n-2,\cdots,1)$$

选主元的 Doolittle 分解法，特别适用于在同一个计算问题中须要求解多个具有相同系数矩阵 $\boldsymbol{A}$ 而具有不同右端向量的线性方程组。此时上述算法的第一部分作分解 $\boldsymbol{QA}=\boldsymbol{LU}$ 只须执行一次，所得的矩阵 $\boldsymbol{L},\boldsymbol{U}$ 和数组 $M(n)$ 适用于每个方程组。不同的方程组只是第二、三两部分的结果不相同。

2.2.3 三角分解法解带状线性方程组

定义 设矩阵 $\boldsymbol{A}=[a_{ij}]_{n\times n}$，如果存在两个正整数 r 和 s，使得当 $i-j>r$ 以及 $j-i>s$ 时，$a_{ij}=0$，即

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & \cdots & a_{1,s+1} & & & \\ \vdots & \ddots & & \ddots & & \\ a_{r+1,1} & & \ddots & & \ddots & \\ & \ddots & & \ddots & & a_{n-s,n} \\ & & \ddots & & \ddots & \vdots \\ & & & a_{n,n-r} & \cdots & a_{nn} \end{bmatrix}$$

则称 $\boldsymbol{A}$ 是上半带宽为 s、下半带宽为 r 的带状矩阵，线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 称为带状线性方程组。

定理 2.5(保带状结构三角分解) 设

(1) $\boldsymbol{A}=[a_{ij}]_{n\times n}$ 是上半带宽为 s、下半带宽为 r 的带状矩阵；

(2) $\boldsymbol{A}$ 的前 $n-1$ 个顺序主子式均不为零。

则 $\boldsymbol{A}$ 有唯一的 Doolittle 分解 $\boldsymbol{A}=\boldsymbol{LU}$，其中 $\boldsymbol{L}$ 是下半带宽为 r 的单位下三角矩阵，$\boldsymbol{U}$ 是上半带宽为 s 的上三角矩阵，即

$$A=LU=\begin{bmatrix}1 & & & & & \\ l_{21} & 1 & & & & \\ \vdots & \ddots & \ddots & & & \\ l_{r+1,1} & & \ddots & \ddots & & \\ & \ddots & & \ddots & \ddots & \\ & & l_{n,n-r} & \cdots & l_{n,n-1} & 1\end{bmatrix}\begin{bmatrix}u_{11} & u_{12} & \cdots & u_{1,s+1} & & \\ & \ddots & & \ddots & \ddots & \\ & & \ddots & & \ddots & u_{n-s,n} \\ & & & \ddots & \ddots & \vdots \\ & & & & \ddots & u_{n-1,n} \\ & & & & & u_{nn}\end{bmatrix}$$

证 由条件(2)并根据定理 2.3,A 必有唯一的 Doolittle 分解 $A=LU$。

当 $i-j>r$ 时,由条件(1)可知 $a_{ik}=0(k=1,2,\cdots,j)$,再由分解计算公式[参见式(2.12)]

$$l_{ik}=(a_{ik}-\sum_{t=1}^{k-1}l_{it}u_{tk})/u_{kk}$$

可推出 $l_{i1}=a_{i1}/u_{11}=0,l_{i2}=(a_{i2}-l_{i1}u_{12})/u_{22}=0,l_{i3}=0,\cdots,l_{i,j-1}=0,l_{ij}=0$。故 L 是下半带宽为 r 的单位下三角矩阵。

当 $j-i>s$ 时,由条件(1)可知 $a_{kj}=0(k=1,2,\cdots,i)$,再由分解计算公式[参见式(2.12)]

$$u_{kj}=a_{kj}-\sum_{t=1}^{k-1}l_{kt}u_{tj}$$

可推出 $u_{1j}=a_{1j}=0,u_{2j}=a_{2j}-l_{21}u_{1j}=0,u_{3j}=0,\cdots,u_{i-1,j}=0,u_{ij}=0$。故 U 是上半带宽为 s 的上三角矩阵。

证毕。

推论 设矩阵 $A=[a_{ij}]_{n\times n}$ 满足定理 2.5 的条件(1)、(2),则 A 有唯一的 Crout 分解 $A=\tilde{L}\tilde{U}$,其中 $\tilde{L}$ 是下半带宽为 r 的下三角矩阵,$\tilde{U}$ 是上半带宽为 s 的单位上三角矩阵。

(证明由读者自己完成)

设 n 元带状线性方程组 $Ax=b$ 的系数矩阵 $A=[a_{ij}]_{n\times n}$ 满足定理 2.5 的条件(1)、(2),今用 Doolittle 分解法求解此方程组。对 A 进行三角分解时,L 的带以下的元素和 U 的带以上的元素不必计算,因而由计算公式(2.12)、(2.13)可得到 Doolittle 分解法求解 n 元带状线性方程组的算法如下:

(1) 作分解 $A=LU$。

对于 $k=1,2,\cdots,n$ 计算

$$u_{kj}=a_{kj}-\sum_{t=\max(1,k-r,j-s)}^{k-1}l_{kt}u_{tj}\quad [j=k,k+1,\cdots,\min(k+s,n)]$$

$$l_{ik}=(a_{ik}-\sum_{t=\max(1,i-r,k-s)}^{k-1}l_{it}u_{tk})/u_{kk}\quad [i=k+1,k+2,\cdots,\min(k+r,n);\ k<n]$$

(2) 求解 $Ly=b,Ux=y$

$$y_1=b_1$$

$$y_i=b_i-\sum_{t=\max(1,i-r)}^{i-1}l_{it}y_t\quad (i=2,3,\cdots,n)$$

$$x_n=y_n/u_{nn}$$

$$x_i=(y_i-\sum_{t=i+1}^{\min(i+s,n)}u_{it}x_t)/u_{ii}\quad (i=n-1,n-2,\cdots,1)$$

对于大型 n 元带状线性方程组 $Ax=b$,当 $n\gg r+s+1$(A 的总带宽)时,为了节省存储量,A

的带外元素不给存储，仅存 $\boldsymbol{A}$ 的带内元素。为此，设置一个二维数组 $\boldsymbol{C}(m,n)$，用于存放 $\boldsymbol{A}$ 的带内元素，其中 $m=r+s+1$。数组 $\boldsymbol{C}$ 的第 j 列存放 $\boldsymbol{A}$ 的第 j 列带内元素，并使 $\boldsymbol{A}$ 的主对角线元素存放在 $\boldsymbol{C}$ 的第 $s+1$ 行中。数组 $\boldsymbol{C}$ 存放完 $\boldsymbol{A}$ 的带内元素后，多出的单元可取零，这些零元素共有 $(1+2+\cdots+s)+(1+2+\cdots+r)$ 个。例如，当矩阵 $\boldsymbol{A}=[a_{ij}]_{6\times6}$ 的下半带宽 $r=1$、上半带宽 $s=2$ 时，应设置数组 $\boldsymbol{C}(4,6)$ 的状况为

$$\boldsymbol{C}=\begin{bmatrix} 0 & 0 & a_{13} & a_{24} & a_{35} & a_{46} \\ 0 & a_{12} & a_{23} & a_{34} & a_{45} & a_{56} \\ a_{11} & a_{22} & a_{33} & a_{44} & a_{55} & a_{66} \\ a_{21} & a_{32} & a_{43} & a_{54} & a_{65} & 0 \end{bmatrix}$$

在数组 $\boldsymbol{C}$ 中检索矩阵 $\boldsymbol{A}$ 的带内元素 a_{ij} 的方法是

$$\boldsymbol{A}\text{ 的带内元素 } a_{ij} = \boldsymbol{C}\text{ 中的元素 } c_{i-j+s+1,j}$$

按照上述对带状矩阵 $\boldsymbol{A}$ 的存储方法和元素 a_{ij} 的检索方法，并且把三角分解的结果 u_{kj} 和 l_{ik} 分别存放在 a_{kj} 和 a_{ik} 原先的存储单元内，那么用 Doolittle 分解法求解 n 元带状线性方程组的算法可重新表述如下(其中“$:=$”表示赋值)：

(1) 作分解 $\boldsymbol{A}=\boldsymbol{LU}$。

对于 $k=1,2,\cdots,n$ 执行

$$c_{k-j+s+1,j} := c_{k-j+s+1,j} - \sum_{t=\max(1,k-r,j-s)}^{k-1} c_{k-t+s+1,t}c_{t-j+s+1,j}$$

$$[j=k,k+1,\cdots,\min(k+s,n)]$$

$$c_{i-k+s+1,k} := \left(c_{i-k+s+1,k} - \sum_{t=\max(1,i-r,k-s)}^{k-1} c_{i-t+s+1,t}c_{t-k+s+1,k}\right)/c_{s+1,k}$$

$$[i=k+1,k+2,\cdots,\min(k+r,n);\ k<n]$$

(2) 求解 $\boldsymbol{Ly}=\boldsymbol{b},\boldsymbol{Ux}=\boldsymbol{y}$(数组 $\boldsymbol{b}$ 先是存放原方程组右端向量，后来存放中间向量 $\boldsymbol{y}$)

$$b_i := b_i - \sum_{t=\max(1,i-r)}^{i-1} c_{i-t+s+1,t}b_t \quad (i=2,3,\cdots,n)$$

$$x_n := b_n/c_{s+1,n}$$

$$x_i := \left(b_i - \sum_{t=i+1}^{\min(i+s,n)} c_{i-t+s+1,t}x_t\right)/c_{s+1,i} \quad (i=n-1,n-2,\cdots,1)$$

2.2.4 追赶法求解三对角线性方程组

设 n 元线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的系数矩阵 $\boldsymbol{A}$ 为非奇异的三对角矩阵

$$\boldsymbol{A}=\begin{bmatrix} a_1 & c_1 & & & \\ d_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & d_n & a_n \end{bmatrix}$$

这种方程组称为三对角线性方程组。显然，$\boldsymbol{A}$ 是上下半带宽都是 1 的带状矩阵。设 $\boldsymbol{A}$ 的前 $n-1$个顺序主子式都不为零，根据定理2.5的推论，$\boldsymbol{A}$ 有唯一的 Crout 分解，并且是保带宽的，即有

$$A=\tilde{L}\tilde{U}=\begin{bmatrix} p_1 & & & & \\ r_2 & p_2 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & r_n & p_n \end{bmatrix}\begin{bmatrix} 1 & q_1 & & & \\ & 1 & q_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & q_{n-1} \\ & & & & 1 \end{bmatrix} \tag{2.18}$$

利用矩阵相乘法，由式(2.18)得

$$a_1=p_1,\quad c_1=p_1q_1$$

对于 $i=1,2,3,\cdots,n$ 有

$$d_i=r_i,\quad a_{i+1}=r_{i+1}q_i+p_{i+1},\quad c_i=p_iq_i(i<n)$$

由此得到三对角矩阵 $\boldsymbol{A}$ 的 Crout 分解计算公式和求解 $\tilde{\boldsymbol{L}}\boldsymbol{y}=\boldsymbol{b}$ 与 $\tilde{\boldsymbol{U}}\boldsymbol{x}=\boldsymbol{y}$ 的计算公式：

$$\begin{cases} p_1=a_1 \\ \text{对于 } i=1,2,\cdots,n-1 \text{ 计算} \\ q_i=c_i/p_i \\ p_{i+1}=a_{i+1}-d_{i+1}q_i \end{cases} \tag{2.19}$$

$$\begin{cases} y_1=b_1/p_1 \\ y_i=(b_i-d_iy_{i-1})/p_i \quad (i=2,3,\cdots,n) \\ x_n=y_n \\ x_i=y_i-q_ix_{i+1} \quad (i=n-1,n-2,\cdots,1) \end{cases} \tag{2.20}$$

式(2.19)和式(2.20)组成的算法称为求解 n 元三对角线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的追赶法。此算法所需的存储量和计算量都很小，只须使用一维数组存放数据，乘除法运算的总次数只有 $5n-4$ 次。

例 3　在四位十进制的限制下，试用追赶法求解下列方程组

$$\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & 1 & 4 & 1 & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}=\begin{bmatrix} 1 \\ 0.5 \\ -1 \\ 3 \\ 2 \end{bmatrix}$$

解　利用计算公式(2.19)，对系数矩阵 $\boldsymbol{A}$ 作 Crout 分解 $\boldsymbol{A}=\tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$，结果为

$$\tilde{L}=\begin{bmatrix} 4.000 & & & & \\ 1 & 3.750 & & & \\ & 1 & 3.733 & & \\ & & 1 & 3.732 & \\ & & & 1 & 3.732 \end{bmatrix}$$

$$\tilde{U}=\begin{bmatrix} 1 & 0.2500 & & & \\ & 1 & 0.2667 & & \\ & & 1 & 0.2679 & \\ & & & 1 & 0.2680 \\ & & & & 1 \end{bmatrix}$$

利用公式(2.20)求出以下结果：

$$y_1=0.2500,\quad y_2=0.06667,\quad y_3=-0.2858$$
$$y_4=0.8805,\quad y_5=0.3001$$
$$x_5=0.3001,\quad x_4=0.8001,\quad x_3=-0.5001$$
$$x_2=0.2001,\quad x_1=0.2000$$

此方程组的精确解为

$$x_5=0.3,\quad x_4=0.8,\quad x_3=-0.5,\quad x_2=0.2,\quad x_1=0.2$$

2.2.5 拟三对角线性方程组的求解方法

系数矩阵为如下的拟三对角矩阵

$$A=\begin{bmatrix} a_1 & c_1 & & & & d_1 \\ d_2 & a_2 & c_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & d_{n-1} & a_{n-1} & c_{n-1} \\ c_n & & & & d_n & a_n \end{bmatrix}$$

的线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 称为拟三对角线性方程组，许多科学技术问题往往最后归结为求解这种线性方程组。容易证明，当拟三对角矩阵 $\boldsymbol{A}$ 满足定理 2.3 的条件时，它可分解为

$$A=\begin{bmatrix} p_1 & & & & & \\ d_2 & p_2 & & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & d_{n-1} & p_{n-1} & \\ r_1 & r_2 & \cdots & r_{n-2} & r_{n-1} & r_n \end{bmatrix}\begin{bmatrix} 1 & q_1 & & & s_1 \\ & 1 & q_2 & & s_2 \\ & & \ddots & \ddots & \vdots \\ & & & \ddots\quad q_{n-2} & s_{n-2} \\ & & & 1 & s_{n-1} \\ & & & & 1 \end{bmatrix}=\boldsymbol{LU}$$

并且有如下计算 p_i,q_i,s_i,r_i 的计算公式：

$$\begin{cases} p_1=a_1 \\ \text{对于 } i=1,2,\cdots,n-2 \text{ 计算} \\ q_i=c_i/p_i \\ p_{i+1}=a_{i+1}-d_{i+1}q_i \end{cases}$$

$$\begin{cases} s_1=d_1/p_1 \\ s_i=-d_is_{i-1}/p_i \quad (i=2,3,\cdots,n-2) \\ s_{n-1}=(c_{n-1}-d_{n-1}s_{n-2})/p_{n-1} \end{cases}$$

$$\begin{cases} r_1=c_n \\ r_j=-r_{j-1}q_{j-1} \quad (j=2,3,\cdots,n-2) \\ r_{n-1}=d_n-r_{n-2}q_{n-2} \\ r_n=a_n-\sum_{j=1}^{n-1}r_js_j \end{cases}$$

于是，由 $\boldsymbol{Ly}=\boldsymbol{b}$ 和 $\boldsymbol{Ux}=\boldsymbol{y}$ 分别解出的 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 为

$$\begin{cases} y_1 = b_1/p_1 \\ y_i = (b_i - d_i y_{i-1})/p_i \quad (i = 2,3,\cdots,n-1) \\ y_n = (b_n - \sum_{j=1}^{n-1} r_j y_j)/r_n \end{cases}$$

$$\begin{cases} x_n = y_n \\ x_{n-1} = y_{n-1} - s_{n-1} x_n \\ x_i = y_i - q_i x_{i+1} - s_i x_n \quad (i = n-2,\cdots,2,1) \end{cases}$$

2.3 矩阵的条件数与病态线性方程组

2.3.1 矩阵的条件数与线性方程组的性态

线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的系数矩阵 $\boldsymbol{A}$ 与右端向量 $\boldsymbol{b}$ 的元素往往是通过观测或计算而得到，因而会带有误差。即使原始数据是精确的，但存放到计算机后由于受字长的限制也会变为近似数。这样，即使解法和计算过程完全精确，也得不到原方程组的精确解。现在要研究，当 $\boldsymbol{A}$ 和 $\boldsymbol{b}$ 分别有了微小变化 $\Delta\boldsymbol{A}$ 和 $\Delta\boldsymbol{b}$ 后，如何影响原方程组解向量 $\boldsymbol{x}$ 的变化 $\Delta\boldsymbol{x}$。

定理 2.6　设 $\boldsymbol{A}$、$\Delta\boldsymbol{A}\in\mathbf{R}^{n\times n}$，$\boldsymbol{b}$、$\Delta\boldsymbol{b}\in\mathbf{R}^n$，$\boldsymbol{A}$ 非奇异，$\boldsymbol{b}\neq\boldsymbol{0}$，$\boldsymbol{x}$ 是方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的解向量。若 $\|\Delta\boldsymbol{A}\|<\dfrac{1}{\|\boldsymbol{A}^{-1}\|}$，则有

(1) 方程组

$$(\boldsymbol{A}+\Delta\boldsymbol{A})(\boldsymbol{x}+\Delta\boldsymbol{x}) = \boldsymbol{b}+\Delta\boldsymbol{b} \tag{2.21}$$

有唯一解 $\boldsymbol{x}+\Delta\boldsymbol{x}$；

(2) 下列估计式成立：

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqslant \frac{\|\boldsymbol{A}\|\ \|\boldsymbol{A}^{-1}\|}{1-\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|}\left(\frac{\|\Delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}+\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right) \tag{2.22}$$

证

(1)
$$\boldsymbol{A}+\Delta\boldsymbol{A}=\boldsymbol{A}(\boldsymbol{I}+\boldsymbol{A}^{-1}\Delta\boldsymbol{A})$$

因 $\|\boldsymbol{A}^{-1}\Delta\boldsymbol{A}\|\leqslant\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|<1$，由定理 1.5 可知 $\boldsymbol{I}+\boldsymbol{A}^{-1}\Delta\boldsymbol{A}$ 非奇异；又因 $\boldsymbol{A}$ 非奇异，故 $\boldsymbol{A}+\Delta\boldsymbol{A}$ 非奇异，方程组(2.21)有唯一解 $\boldsymbol{x}+\Delta\boldsymbol{x}$。

(2) 由方程(2.21)，并注意到 $\boldsymbol{b}=\boldsymbol{Ax}$ 以及 $\|\boldsymbol{b}\|\leqslant\|\boldsymbol{A}\|\ \|\boldsymbol{x}\|$，$\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|<1$，可得

$$\boldsymbol{A}\Delta\boldsymbol{x}+\Delta\boldsymbol{A}\boldsymbol{x}+\Delta\boldsymbol{A}\Delta\boldsymbol{x}=\Delta\boldsymbol{b}$$

$$\Delta\boldsymbol{x}=\boldsymbol{A}^{-1}\Delta\boldsymbol{b}-\boldsymbol{A}^{-1}\Delta\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}^{-1}\Delta\boldsymbol{A}\Delta\boldsymbol{x}$$

$$\|\Delta\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{b}\|+\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|\ \|\boldsymbol{x}\|+\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|\ \|\Delta\boldsymbol{x}\|$$

$$(1-\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|)\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\|\boldsymbol{A}\|\ \|\boldsymbol{A}^{-1}\|\ \frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}+\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|$$

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\|\boldsymbol{A}\|\ \|\boldsymbol{A}^{-1}\|}{1-\|\boldsymbol{A}^{-1}\|\ \|\Delta\boldsymbol{A}\|}\left(\frac{\|\Delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}+\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right)$$

证毕。

把式(2.22)写成

$$\frac{\|\Delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqslant \frac{\|\boldsymbol{A}\| \ \|\boldsymbol{A}^{-1}\|}{1-\|\boldsymbol{A}\| \ \|\boldsymbol{A}^{-1}\| \ \frac{\|\Delta \boldsymbol{A}\|}{\|\boldsymbol{A}\|}}\left(\frac{\|\Delta \boldsymbol{A}\|}{\|\boldsymbol{A}\|}+\frac{\|\Delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right)$$

由此看出，量 $\|\boldsymbol{A}\| \ \|\boldsymbol{A}^{-1}\|$ 越小，系数矩阵 $\boldsymbol{A}$ 和右端向量 $\boldsymbol{b}$ 的相对误差对解向量的相对误差的影响就越小；反之，则影响可能越大。

定义　对非奇异矩阵 $\boldsymbol{A}$，称量 $\|\boldsymbol{A}\| \ \|\boldsymbol{A}^{-1}\|$ 为矩阵 $\boldsymbol{A}$ 的条件数，记作

$$\operatorname{cond}(\boldsymbol{A})=\|\boldsymbol{A}\| \ \|\boldsymbol{A}^{-1}\|$$

矩阵 $\boldsymbol{A}$ 的条件数与所取的矩阵范数有关，常用的条件数是

$$\operatorname{cond}(\boldsymbol{A})_{\infty}=\|\boldsymbol{A}\|_{\infty}\|\boldsymbol{A}^{-1}\|_{\infty}, \quad \operatorname{cond}(\boldsymbol{A})_{2}=\|\boldsymbol{A}\|_{2}\|\boldsymbol{A}^{-1}\|_{2}$$

矩阵 $\boldsymbol{A}$ 的条件数有以下性质：

(1) 对任何非奇异矩阵 $\boldsymbol{A}$，$\operatorname{cond}(\boldsymbol{A})\geqslant 1$。

(2) 设 $\boldsymbol{A}$ 是非奇异矩阵，$k\neq 0$ 是常数，则有 $\operatorname{cond}(k\boldsymbol{A})=\operatorname{cond}(\boldsymbol{A})$。

(3) 设 $\boldsymbol{A}$ 是非奇异的实对称矩阵，则有

$$\operatorname{cond}(\boldsymbol{A})_{2}=\left|\frac{\lambda_{1}}{\lambda_{n}}\right|$$

其中 λ_1 和 λ_n 分别是矩阵 $\boldsymbol{A}$ 的模为最大和模为最小的特征值。

(4) 设 $\boldsymbol{A}$ 是正交矩阵，则有 $\operatorname{cond}(\boldsymbol{A})_2=1$。

定义　设线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的系数矩阵 $\boldsymbol{A}$ 非奇异，若 $\operatorname{cond}(\boldsymbol{A})$ 相对很大，则称 $\boldsymbol{Ax}=\boldsymbol{b}$ 是病态线性方程组(也称 $\boldsymbol{A}$ 是病态矩阵)；若 $\operatorname{cond}(\boldsymbol{A})$ 相对较小，则称 $\boldsymbol{Ax}=\boldsymbol{b}$ 是良态线性方程组(也称 $\boldsymbol{A}$ 是良态矩阵)。

矩阵 $\boldsymbol{A}$ 的条件数刻画了线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的性态。$\boldsymbol{A}$ 的条件数越大，方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的病态程度越严重。对于严重病态的线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，当 $\boldsymbol{A}$ 和 $\boldsymbol{b}$ 有微小变化时，即使求解过程是精确进行的，所得的解相对于原方程组的解也会有很大的相对误差。

例 4　设有线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 为

$$\begin{bmatrix} 1 & 1.0001 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

(1) 求 $\operatorname{cond}(\boldsymbol{A})_{\infty}$ 和 $\boldsymbol{Ax}=\boldsymbol{b}$ 的解 $\boldsymbol{x}$。

(2) 设 $\boldsymbol{b}$ 变化为 $\boldsymbol{b}+\Delta\boldsymbol{b}=(2.0001,2)^{\mathrm{T}}$，求 $\boldsymbol{A}(\boldsymbol{x}+\Delta\boldsymbol{x})=\boldsymbol{b}+\Delta\boldsymbol{b}$ 的解 $\boldsymbol{x}+\Delta\boldsymbol{x}$。

(3) 计算 $\dfrac{\|\Delta\boldsymbol{b}\|_{\infty}}{\|\boldsymbol{b}\|_{\infty}}$ 和 $\dfrac{\|\Delta\boldsymbol{x}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}}$。

解　(1)

$$\boldsymbol{A}^{-1}=\begin{bmatrix} -10^{4} & 1.0001\times 10^{4} \\ 10^{4} & -10^{4} \end{bmatrix}$$

$$\operatorname{cond}(\boldsymbol{A})_{\infty}=\|\boldsymbol{A}\|_{\infty}\|\boldsymbol{A}^{-1}\|_{\infty}=2.0001\times(2.0001\times 10^{4})\approx 4\times 10^{4}$$

$\boldsymbol{Ax}=\boldsymbol{b}$ 的解为 $\boldsymbol{x}=(2,0)^{\mathrm{T}}$；

(2) $\boldsymbol{A}(\boldsymbol{x}+\Delta\boldsymbol{x})=\boldsymbol{b}+\Delta\boldsymbol{b}$ 的解为 $\boldsymbol{x}+\Delta\boldsymbol{x}=(1,1)^{\mathrm{T}}$；

(3) $\Delta\boldsymbol{b}=(0.0001,0)^{\mathrm{T}}$，$\Delta\boldsymbol{x}=(-1,1)^{\mathrm{T}}$，

$$\frac{\|\Delta\boldsymbol{b}\|_{\infty}}{\|\boldsymbol{b}\|_{\infty}}=0.005\%, \quad \frac{\|\Delta\boldsymbol{x}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}}=50\%$$

从例 4 可看出，由右端向量 $\boldsymbol{b}$ 的微小相对误差 0.005% 引起解的很大相对误差 50%，后者是前者的 10^4 倍。原因就是矩阵 $\boldsymbol{A}$ 的条件数很大，方程组病态比较严重。

对于严重病态的线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，即使原始数据 $\boldsymbol{A}$ 和 $\boldsymbol{b}$ 没有误差，但在求解过程中只要有舍入误差，所得的解也会有很大的相对误差。

例 5　设线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 为

$$\begin{bmatrix} 0.2161 & 0.1441 \\ 1.2969 & 0.8648 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 0.144 \\ 0.8642 \end{bmatrix}$$

(1) 求 $\operatorname{cond}(\boldsymbol{A})_\infty$；

(2) 在八位十进制的限制下，用列主元素 Gauss 消去法求解。

解　(1)

$$\boldsymbol{A}^{-1}=\begin{bmatrix} -0.8648\times 10^8 & 0.1441\times 10^8 \\ 1.2969\times 10^8 & -0.2161\times 10^8 \end{bmatrix}$$

$$\operatorname{cond}(\boldsymbol{A})_\infty=\|\boldsymbol{A}\|_\infty\|\boldsymbol{A}^{-1}\|_\infty\approx 3.27\times 10^8$$

(2)

$$\begin{bmatrix} 0.2161 & 0.1441 & 0.1440 \\ 1.2969 & 0.8648 & 0.8642 \end{bmatrix}\rightarrow$$

$$\begin{bmatrix} 1.2969 & 0.8648 & 0.8642 \\ 0.2161 & 0.1441 & 0.1440 \end{bmatrix}\rightarrow$$

$$\begin{bmatrix} 1.2969 & 0.8648 & 0.8642 \\ 0 & 1\times 10^{-8} & -1\times 10^{-8} \end{bmatrix}$$

得到解 $x_2=-1$，$x_1=1.3331791$。

例 5 第(2)问所得的解与原方程组的精确解 $x_1=2$，$x_2=-2$ 相比较，其相对误差为

$$\frac{\|\Delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}=50\%$$

在八位十进制的限制下，每运算一次的舍入误差并不大，但所得到的解的精确性却很低，原因就是该方程组系数矩阵的条件数很大，方程组病态很严重。

2.3.2 关于病态线性方程组的求解问题

可以用下列的方法判别线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 是否病态：

(1) 当 $|\det\boldsymbol{A}|$ 相对很小或 $\boldsymbol{A}$ 的某些行(或列)近似线性相关时，方程组可能病态。

(2) 用列主元素 Gauss 消去法求解方程组时，若出现小列主元 $|a_{i_k k}^{(k)}|\ll 1$，则方程组可能病态。

(3) 分别用 $\boldsymbol{b}$ 和 $\boldsymbol{b}+\Delta\boldsymbol{b}$($\|\Delta\boldsymbol{b}\|\ll 1$)作方程组的右端向量，求解 $\boldsymbol{Ax}=\boldsymbol{b}$ 和 $\boldsymbol{A}\tilde{\boldsymbol{x}}=\boldsymbol{b}+\Delta\boldsymbol{b}$，若 $\boldsymbol{x}$ 和 $\tilde{\boldsymbol{x}}$ 相差很大，则 $\boldsymbol{Ax}=\boldsymbol{b}$ 是病态的。

(4) 当 $\boldsymbol{A}$ 的元素的数量级差别很大，且无一定规则时，方程组可能病态。例如，

$$\boldsymbol{A}=\begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 10^{10} \end{bmatrix},\quad \operatorname{cond}(\boldsymbol{A})_\infty=\frac{(10^{10}+0.1)^2}{10^9-0.01}\approx 10^{11}$$

相应的方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 是病态的。但对于

$$B = \begin{bmatrix} 10^{10} & 0.1 \\ 0.1 & 10^{10} \end{bmatrix},\quad \text{cond}(B)_\infty = \frac{(10^{10}+0.1)^2}{10^{20}-0.01} \approx 1$$

相应的方程组 $Bx=b$ 是良态的。

对于病态线性方程组可采用以下的方法求解：

(1) 采用高精度的算术运算，例如采用双精度，可改善和减轻病态矩阵的影响。但有时还不行。

(2) 平衡方法。当 A 的元素的数量级差别很大时，采用行平衡或列平衡的方法可降低 A 的条件数。

设 n 元方程组 $Ax=b$，$A=[a_{ij}]_{n\times n}$ 非奇异。所谓行平衡方法就是：计算 $s_i=\max\limits_{1\leqslant j\leqslant n}|a_{ij}|$ $(i=1,2,\cdots,n)$，令

$$D = \text{diag}\left(\frac{1}{s_1},\frac{1}{s_2},\cdots,\frac{1}{s_n}\right)$$

得到与原方程组同解的方程组 $DAx=Db$，此方程组的系数矩阵 DA 的条件数有可能大大低于 A 的条件数。例如，原方程组 $Ax=b$ 为

$$\begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 10^{10} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 10^{10} \end{bmatrix}$$

因 A 的条件数 $\text{cond}(A)_\infty \approx 10^{11}$ 很大，故此方程组是病态的。进行行平衡：$s_1=0.1$，$s_2=10^{10}$，$D=\text{diag}(10,10^{-10})$，得到同解方程组 $DAx=Db$ 为

$$\begin{bmatrix} 1 & 1 \\ 10^{-11} & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

此时，DA 的条件数 $\text{cond}(DA)_\infty = \dfrac{4}{1-10^{-11}} \approx 4$ 很小，方程组 $DAx=Db$ 是良态的。

(3) 残差校正法。设 A 非奇异且 $\text{cond}(A)$ 不特别大，方程组 $Ax=b$ 病态但不特别严重，这时可用残差校正法求解 $Ax=b$。

先使用通常的方法在计算机上求解 $Ax=b$，得到 $Ax=b$ 的近似解 $\tilde{x}$。为了校正 $\tilde{x}$，计算残差

$$r = b - A\tilde{x}$$

并求解方程组 $A\Delta x=r$，得到 $\tilde{x}$ 的修正量 Δx，令

$$\tilde{\tilde{x}} = \tilde{x} + \Delta x$$

则 $\tilde{\tilde{x}}$ 就是经过校正的解向量。如果计算残差 r、求解 Δx 以及计算 $\tilde{\tilde{x}}$ 都是精确的，那么，就有

$$A\tilde{\tilde{x}} = A(\tilde{x}+\Delta x) = (b-r)+r = b$$

即 $\tilde{\tilde{x}}$ 就是原方程组的精确解。但实际计算时总有舍入误差，特别是求解 Δx，得到的是 $A\Delta x=r$ 的近似解。所以，应重复执行上述过程，直至满足精度要求为止。这就是残差校正法。此法执行过程中，残差的计算应尽量精确，以保证得到有效的修正量 Δx。残差校正法求解 $Ax=b$ 的具体算法如下：

(1) 对 A 进行选主元的 Doolittle 分解

$$QA = LU$$

(2) 求解 $Ly=Qb$ 和 $Ux^{(1)}=y$ 得初始解向量 $x^{(1)}$。

(3) 对于 $k=1,2,\cdots,M$ 执行

① 计算残差 $\boldsymbol{r}^{(k)}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(k)}$ (用双精度)。

② 求解 $\boldsymbol{L}\boldsymbol{y}^{(k)}=\boldsymbol{Q}\boldsymbol{r}^{(k)}$ 和 $\boldsymbol{U}\boldsymbol{d}^{(k)}=\boldsymbol{y}^{(k)}$。

③ 如果 $\|\boldsymbol{d}^{(k)}\|_\infty/\|\boldsymbol{x}^{(k)}\|_\infty\leqslant 10^{-t}$,则停止计算,$\boldsymbol{x}^{(k)}$ 就作为 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的解;否则转④。

④ 计算 $\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\boldsymbol{d}^{(k)}$。

(4) 输出 M 次校正失败的信息。

算法中的 t 可取计算机所能达到的十进制浮点数的位数。

2.4 迭代法

迭代法求解方程组(2.2),就是构造一个无限的向量序列,使它的极限是方程组(2.2)的解向量。即使计算过程是精确进行的,迭代法也不能通过有限次算术运算求得方程组(2.2)的精确解,而只能逐步逼近它。因此,凡是迭代法都存在收敛性与精度控制的问题。

迭代法常用于求解大型稀疏线性方程组。

2.4.1 迭代法的一般形式及其收敛性

设 n 元线性方程组(2.2),即

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$$

的系数矩阵 $\boldsymbol{A}=[a_{ij}]\in\mathbf{R}^{n\times n}$ 非奇异,右端向量 $\boldsymbol{b}\neq\boldsymbol{0}$,因而方程组(2.2)有唯一的非零解向量。

把系数矩阵 $\boldsymbol{A}$ 分裂成两个矩阵 $\boldsymbol{N}$ 和 $\boldsymbol{P}$ 的差

$$\boldsymbol{A}=\boldsymbol{N}-\boldsymbol{P}$$

其中 $\boldsymbol{N}$ 是非奇异的,代入式(2.2),得

$$\boldsymbol{N}\boldsymbol{x}=\boldsymbol{P}\boldsymbol{x}+\boldsymbol{b}$$

$$\boldsymbol{x}=\boldsymbol{N}^{-1}\boldsymbol{P}\boldsymbol{x}+\boldsymbol{N}^{-1}\boldsymbol{b}$$

记

$$\boldsymbol{G}=\boldsymbol{N}^{-1}\boldsymbol{P},\quad \boldsymbol{d}=\boldsymbol{N}^{-1}\boldsymbol{b}$$

则由上式得到

$$\boldsymbol{x}=\boldsymbol{G}\boldsymbol{x}+\boldsymbol{d} \tag{2.23}$$

方程组(2.23)与原方程组(2.2)是同解方程组。

任取一个向量 $\boldsymbol{x}^{(0)}\in\mathbf{R}^{n\times n}$ 作为方程组(2.2)的初始近似解,按递推公式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{G}\boldsymbol{x}^{(k)}+\boldsymbol{d}\quad(k=0,1,\cdots) \tag{2.24}$$

产生一个向量序列 $\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\cdots,\boldsymbol{x}^{(k)},\cdots$,当 k 足够大时,以此序列中的向量 $\boldsymbol{x}^{(k)}$ 作为方程组(2.2)的近似解。这种求解方程组(2.2)的方法称为迭代法,递推公式(2.24)称为迭代公式,其中的矩阵 $\boldsymbol{G}$ 称为迭代矩阵。公式(2.24)就是求解方程组(2.2)的迭代法(严格地说,是线性迭代法)的一般形式。

为了研究迭代公式(2.24)的收敛性,首先介绍向量序列收敛的概念。

定义 设有向量序列

$$\boldsymbol{x}^{(k)}=(x_1^{(k)},x_2^{(k)},\cdots,x_n^{(k)})^{\mathrm{T}}\quad(k=0,1,\cdots)$$

如果存在常向量 $\boldsymbol{x}^*=(x_1^*,x_2^*,\cdots,x_n^*)^{\mathrm{T}}$,使得

$$\lim_{k\to\infty} x_i^{(k)} = x_i^* \quad (i = 1,2,\cdots,n)$$

则称向量序列$\{\boldsymbol{x}^{(k)}\}$收敛于常向量$\boldsymbol{x}^*$，记为

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*$$

定理 2.7 设有向量序列$\{\boldsymbol{x}^{(k)}\}$和常向量$\boldsymbol{x}^*$，如果对某种范数，有

$$\lim_{k\to\infty} \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \| = 0$$

则必有

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*$$

证 根据向量范数等价性定理 1.2，必有

$$\lim_{k\to\infty} \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \|_\infty = 0$$

即

$$\lim_{k\to\infty} \max_{1\leqslant i\leqslant n} | x_i^{(k)} - x_i^* | = 0$$

因而有

$$\lim_{k\to\infty} x_i^{(k)} = x_i^* \quad (i = 1,2,\cdots,n)$$

根据定义，上式即$\lim\limits_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*$。

证毕。

如果对任取的$\boldsymbol{x}^{(0)}$迭代公式(2.24)所产生的向量序列$\{\boldsymbol{x}^{(k)}\}$都收敛于某一常向量$\boldsymbol{x}^*$，则称该迭代法是收敛的。显然，$\boldsymbol{x}^*$满足方程组(2.23)，即

$$\boldsymbol{x}^* = \boldsymbol{G}\boldsymbol{x}^* + \boldsymbol{d} \tag{2.25}$$

因而$\boldsymbol{x}^*$就是线性方程组(2.2)的解。

只有在迭代法收敛的情况下，用它所产生的向量序列$\{\boldsymbol{x}^{(k)}\}$中的向量作为方程组(2.2)的近似解才有意义，而且，k越大，$\boldsymbol{x}^{(k)}$作为方程组的解就越精确。

定义 设$n\times n$矩阵$\boldsymbol{G}$的特征值是$\lambda_1,\lambda_2,\cdots,\lambda_n$，称

$$\rho(\boldsymbol{G}) = \max_{1\leqslant i\leqslant n} | \lambda_i |$$

为矩阵$\boldsymbol{G}$的谱半径。

定理 2.8 对任意的向量$\boldsymbol{d}$，迭代法(2.24)收敛的充分必要条件是$\rho(\boldsymbol{G})<1$。

证明从略，读者可参阅文献[4]第 151～152 页。

用迭代矩阵的谱半径判断迭代公式(2.24)是否收敛往往不容易，下面给出一个容易使用的判断收敛的充分条件。

定理 2.9 如果矩阵$\boldsymbol{G}$的某种范数$\|\boldsymbol{G}\|<1$，则

(1) 方程组(2.23)的解$\boldsymbol{x}^*$存在且唯一；

(2) 对于迭代公式(2.24)，有

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*, \quad \forall\, \boldsymbol{x}^{(0)} \in \mathbf{R}^n$$

并且下列两式成立

$$\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \| \leqslant \frac{\|\boldsymbol{G}\|^k}{1-\|\boldsymbol{G}\|} \| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \| \tag{2.26}$$

$$\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \| \leqslant \frac{\|\boldsymbol{G}\|}{1-\|\boldsymbol{G}\|} \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \| \tag{2.27}$$

证 (1) 因$\|\boldsymbol{G}\|<1$，根据定理 1.5 可知矩阵$\boldsymbol{I}-\boldsymbol{G}$非奇异，其中$\boldsymbol{I}$是单位矩阵，故方程

组(2.23)的解 x^* 存在且唯一。

(2) 由式(2.24)减去式(2.25)，得

$$x^{(k+1)} - x^* = G(x^{(k)} - x^*)$$

由此得

$$\begin{aligned} 0 \leqslant \| x^{(k+1)} - x^* \| &\leqslant \| G \| \| x^{(k)} - x^* \| \leqslant \\ &\| G \|^2 \| x^{(k-1)} - x^* \| \leqslant \cdots \leqslant \\ &\| G \|^{k+1} \| x^{(0)} - x^* \| \end{aligned}$$

因为 $\| G \| < 1$，所以由上式得

$$\lim_{k\to\infty} \| x^{(k-1)} - x^* \| = 0$$

根据定理 2.7，可知 $\lim\limits_{k\to\infty} x^{(k)} = x^*$ 成立。

设 $m > k$，则有

$$x^{(k)} - x^{(m)} = \sum_{i=k}^{m-1} (x^{(i)} - x^{(i+1)})$$

$$\begin{aligned} \| x^{(k)} - x^{(m)} \| &\leqslant \sum_{i=k}^{m-1} \| x^{(i)} - x^{(i+1)} \| \leqslant \\ &\sum_{i=k}^{m-1} \| G \|^i \| x^{(0)} - x^{(1)} \| = \\ &\| G \|^k \frac{1 - \| G \|^{m-k}}{1 - \| G \|} \| x^{(0)} - x^{(1)} \| \end{aligned}$$

令 $m \to \infty$，由于 $\| G \| < 1$，故由上式得

$$\| x^{(k)} - x^* \| \leqslant \frac{\| G \|^k}{1 - \| G \|} \| x^{(1)} - x^{(0)} \|$$

仍设 $m > k$，则有

$$x^{(k)} - x^{(m)} = \sum_{i=1}^{m-k} (x^{(k+i-1)} - x^{(k+i)})$$

$$\begin{aligned} \| x^{(k)} - x^{(m)} \| &\leqslant \sum_{i=1}^{m-k} \| G \|^i \| x^{(k-1)} - x^{(k)} \| = \\ &\| G \| \frac{1 - \| G \|^{m-k}}{1 - \| G \|} \| x^{(k-1)} - x^{(k)} \| \end{aligned}$$

令 $m \to \infty$，由上式得

$$\| x^{(k)} - x^* \| \leqslant \frac{\| G \|}{1 - \| G \|} \| x^{(k)} - x^{(k-1)} \|$$

证毕。

根据式(2.26)，$\| G \|$ 越小，$x^{(k)}$ 收敛得越快，而且式(2.26)可以作为误差估计式。

根据式(2.27)，当 $\| G \|$ 不是很接近 1 时，只要 $\| x^{(k)} - x^{(k-1)} \|$ 很小，$x^{(k)}$ 就很接近 x^*。所以，在实际应用中，可预先给定一个小的正数 ε，当满足

$$\| x^{(k)} - x^{(k-1)} \| < \varepsilon$$

或

$$\frac{\| x^{(k)} - x^{(k-1)} \|}{\| x^{(k)} \|} < \varepsilon$$

时停止迭代，用当前的 $\boldsymbol{x}^{(k)}$ 作为方程组(2.23)的近似解。

但是，如果 $\|\boldsymbol{G}\|$ 很接近 1，那么即使 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|$ 很小，也不能断定 $\boldsymbol{x}^{(k)}$ 很接近 $\boldsymbol{x}^*$。此外，由于计算机上舍入误差的影响，不要以为只要 $\|\boldsymbol{G}\|<1$ 就能在迭代过程中使 $\boldsymbol{x}^{(k)}$ 可以任意接近 $\boldsymbol{x}^*$。如果方程组(2.23)是病态的，即矩阵 $\boldsymbol{I}-\boldsymbol{G}$ 的条件数很大，那么即使迭代次数大量增加，也未必能得到好的结果。

2.4.2 Jacobi 迭代法

设方程组(2.2)的系数矩阵 $\boldsymbol{A}=[a_{ij}]\in\mathbf{R}^{n\times n}$ 满足条件 $a_{ii}\neq 0(i=1,2,\cdots,n)$。把 $\boldsymbol{A}$ 分解为

$$\boldsymbol{A}=\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{U}$$

这里

$$\boldsymbol{D}=\begin{bmatrix}a_{11}&&&\\&a_{22}&&\\&&\ddots&\\&&&a_{nn}\end{bmatrix},\quad \boldsymbol{L}=\begin{bmatrix}0&&&\\a_{21}&0&&\\\vdots&\ddots&\ddots&\\a_{n1}&\cdots&a_{n,n-1}&0\end{bmatrix},\quad \boldsymbol{U}=\begin{bmatrix}0&a_{12}&\cdots&a_{1n}\\&0&\ddots&\vdots\\&&\ddots&a_{n-1,n}\\&&&0\end{bmatrix}$$

根据已知条件，$\boldsymbol{D}^{-1}$ 存在。

在迭代法一般形式(2.24)中，取 $\boldsymbol{N}=\boldsymbol{D}$，$\boldsymbol{P}=-(\boldsymbol{L}+\boldsymbol{U})$，形成以下的迭代公式

$$\boldsymbol{x}^{(k+1)}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x}^{(k)}+\boldsymbol{D}^{-1}\boldsymbol{b}\quad(k=0,1,\cdots)\tag{2.28}$$

其中 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$ 任取。由迭代公式(2.28)所表示的迭代法称为Jacobi(雅可比)迭代法，又称简单迭代法，它的迭代矩阵是

$$\boldsymbol{G}_{\mathrm{J}}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$$

因 $\boldsymbol{D}^{-1}=\mathrm{diag}(a_{ii}^{-1})$，故 Jacobi 迭代法式(2.28)的分量形式是

$$x_i^{(k+1)}=\Big(-\sum_{\substack{j=1\\j\neq i}}^{n}a_{ij}x_j^{(k)}+b_i\Big)/a_{ii}\quad(i=1,2,\cdots,n;\ k=0,1,\cdots)\tag{2.29}$$

由定理 2.8 得到

定理 2.10 Jacobi 迭代法收敛的充分必要条件是 $\rho(\boldsymbol{G}_{\mathrm{J}})<1$。

由定理 2.9 得到

定理 2.11 如果 $\|\boldsymbol{G}_{\mathrm{J}}\|<1$，则 Jacobi 迭代法收敛。

下面再给出一个收敛的充分条件，它直接使用原方程组(2.2)的系数矩阵进行判断。

引理 2.1 若矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 是主对角线按行(或按列)严格占优阵，则 $\boldsymbol{A}$ 是非奇异矩阵。

证 设 $\boldsymbol{A}=[a_{ij}]\in\mathbf{R}^{n\times n}$ 是主对角线按行严格占优阵，即满足条件

$$|a_{ii}|>\sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}|\quad(i=1,2,\cdots,n)$$

因而 $a_{ii}\neq 0(i=1,2,\cdots,n)$，$\boldsymbol{D}=\mathrm{diag}(a_{ii})$ 非奇异。令

$$\boldsymbol{A}=\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{U}=\boldsymbol{D}[\boldsymbol{I}+\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})]=\boldsymbol{D}(\boldsymbol{I}-\boldsymbol{G})$$

其中 $\boldsymbol{G}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$。由 $\boldsymbol{A}$ 所满足的条件可知

$$\|\boldsymbol{G}\|_\infty=\max_{1\leqslant i\leqslant n}\sum_{\substack{j=1\\j\neq i}}^{n}\left|\frac{a_{ij}}{a_{ii}}\right|<1$$

又由定理 1.5 可知 $\boldsymbol{I}-\boldsymbol{G}$ 非奇异，因而 $\boldsymbol{A}$ 也非奇异。

设 $\boldsymbol{A}$ 是主对角线按列严格占优阵，则 $\boldsymbol{A}^{\mathrm{T}}$ 是主对角线按行严格占优阵，根据前面所证，$\boldsymbol{A}^{\mathrm{T}}$ 非奇异，因此 $\boldsymbol{A}$ 也非奇异。

证毕。

定理 2.12　如果方程组(2.2)的系数矩阵是主对角线按行(或按列)严格占优阵，则用 Jacobi迭代法求解必收敛。

证　因 $\boldsymbol{A}$ 是主对角线按行(或按列)严格占优阵，故 $\boldsymbol{D}^{-1}=\mathrm{diag}(a_{ii}^{-1})$存在。假定 Jacobi 迭代法不收敛，则根据定理 2.8，迭代矩阵 $\boldsymbol{G}_{\mathrm{J}}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$有一特征值 μ 满足 $|\mu|\geqslant 1$，且有

$$\det(\mu\boldsymbol{I}-\boldsymbol{G}_{\mathrm{J}})=0$$

而

$$\begin{aligned}\det(\mu\boldsymbol{I}-\boldsymbol{G}_{\mathrm{J}}) &= \det[\mu\boldsymbol{I}+\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})]=\\ &\det\{\mu\boldsymbol{D}^{-1}[\boldsymbol{D}+\mu^{-1}(\boldsymbol{L}+\boldsymbol{U})]\}=\\ &\mu^{n}\cdot\det\boldsymbol{D}^{-1}\cdot\det[\boldsymbol{D}+\mu^{-1}(\boldsymbol{L}+\boldsymbol{U})]\end{aligned}$$

显然 $\mu^{n}\neq 0$，$\det\boldsymbol{D}^{-1}\neq 0$。又因 $|\mu^{-1}|\leqslant 1$，所以，当 $\boldsymbol{A}$ 是主对角线按行严格占优阵时，有

$$|a_{ii}|>\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\geqslant\sum_{\substack{j=1\\ j\neq i}}^{n}|\mu^{-1}a_{ij}|\quad(i=1,2,\cdots,n)$$

当 $\boldsymbol{A}$ 是主对角线按列严格占优阵时，有

$$|a_{jj}|>\sum_{\substack{i=1\\ i\neq j}}^{n}|a_{ij}|\geqslant\sum_{\substack{i=1\\ i\neq j}}^{n}|\mu^{-1}a_{ij}|\quad(j=1,2,\cdots,n)$$

即，矩阵 $\boldsymbol{D}+\mu^{-1}(\boldsymbol{L}+\boldsymbol{U})$是主对角线按行(或按列)严格占优阵。根据引理 2.1 可知

$$\det[\boldsymbol{D}+\mu^{-1}(\boldsymbol{L}+\boldsymbol{U})]\neq 0$$

于是有

$$\det(\mu\boldsymbol{I}-\boldsymbol{G}_{\mathrm{J}})\neq 0$$

所出现的矛盾证实 Jacobi 迭代法必收敛。

证毕。

例 6　证明用 Jacobi 迭代法求解下列方程组

$$\begin{cases}10x_1-2x_2-2x_3=1\\ -2x_1+10x_2-x_3=0.5\\ -x_1-2x_2+3x_3=1\end{cases}$$

必收敛，并求解，要求 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_{\infty}\leqslant 10^{-5}$。

解　迭代矩阵为

$$\boldsymbol{G}_{\mathrm{J}}=\begin{bmatrix}0 & 0.2 & 0.2\\ 0.2 & 0 & 0.1\\ \dfrac{1}{3} & \dfrac{2}{3} & 0\end{bmatrix}$$

因 $\|\boldsymbol{G}_{\mathrm{J}}\|_{1}=0.2+\dfrac{2}{3}=\dfrac{13}{15}<1$，故 Jacobi 迭代法必收敛，迭代公式为

$$\begin{cases} x_1^{(k+1)} = +0.2x_2^{(k)} + 0.2x_3^{(k)} + 0.1 \\ x_2^{(k+1)} = 0.2x_1^{(k)} + 0.1x_3^{(k)} + 0.05 \\ x_3^{(k+1)} = \dfrac{1}{3}x_1^{(k)} + \dfrac{2}{3}x_2^{(k)} + \dfrac{1}{3} \\ \qquad [(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^{\mathrm{T}} \text{ 任选}; k = 0, 1, \cdots] \end{cases}$$

计算结果见表 2-1。由于 $\|\boldsymbol{x}^{(15)} - \boldsymbol{x}^{(14)}\|_\infty = 10^{-5}$，故所求的解为

$$x_1^* \approx 0.231\,087,\quad x_2^* \approx 0.147\,055,\quad x_3^* \approx 0.508\,393$$

表 2-1 例 6 计算结果

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.0	0.0	0.0
1	0.100 000	0.050 000	0.333 333
2	0.176 667	0.103 333	0.400 000
3	0.200 667	0.125 333	0.461 111
4	0.217 289	0.136 245	0.483 777
5	0.224 004	0.141 836	0.496 593
⋮	⋮	⋮	⋮
13	0.231 069	0.147 041	0.508 362
14	0.231 081	0.147 050	0.508 383
15	0.231 087	0.147 055	0.508 393

例 7 证明：方程组

$$\begin{cases} 2x_1 - x_2 + x_3 = 1 \\ x_1 + x_2 + x_3 = 1 \\ x_1 + x_2 - 2x_3 = 1 \end{cases}$$

使用 Jacobi 迭代法求解不收敛。

证 对于此方程组，Jacobi 迭代法的迭代矩阵为

$$\boldsymbol{G}_{\mathrm{J}} = \begin{bmatrix} 0 & 0.5 & -0.5 \\ -1 & 0 & -1 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

$\boldsymbol{G}_{\mathrm{J}}$ 的特征多项式为

$$\det(\lambda \boldsymbol{I} - \boldsymbol{G}_{\mathrm{J}}) = \begin{vmatrix} \lambda & -0.5 & 0.5 \\ 1 & \lambda & 1 \\ -0.5 & -0.5 & \lambda \end{vmatrix} = \lambda(\lambda^2 + 1.25)$$

$\boldsymbol{G}_{\mathrm{J}}$ 的特征值为 $\lambda_1 = 0, \lambda_2 = \sqrt{1.25}\,\mathrm{i}, \lambda_3 = -\sqrt{1.25}\,\mathrm{i}$，故 $\rho(\boldsymbol{G}_{\mathrm{J}}) = \sqrt{1.25} > 1$，因而 Jacobi 迭代法不收敛。

证毕。

例 8 对方程组

$$\begin{cases} 10x_1 + 2x_2 + \quad\quad 5x_4 = 2 \\ x_1 + 9x_2 + 2x_3 - 4x_4 = 0 \\ 3x_1 - 4x_2 + 8x_3 \quad\quad = 1 \\ x_1 + 2x_2 - x_3 + 6x_4 = 8 \end{cases}$$

使用 Jacobi 迭代法求解，试判断是否收敛？

解　由于此方程组的系数矩阵是主对角线按行严格占优阵，所以，Jacobi 迭代法收敛。

2.4.3 Gauss - Seidel 迭代法

仍设方程组(2.2)的系数矩阵 $\boldsymbol{A}=[a_{ij}]\in\mathbf{R}^{n\times n}$ 满足条件 $a_{ii}\neq 0(i=1,2,\cdots,n)$，$\boldsymbol{A}=\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{U}$。

在迭代法一般形式(2.24)中，取 $\boldsymbol{N}=\boldsymbol{D}+\boldsymbol{L}$，$\boldsymbol{P}=-\boldsymbol{U}$，形成以下的迭代公式

$$\boldsymbol{x}^{(k+1)} = -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{U}\boldsymbol{x}^{(k)} + (\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{b} \quad (k=0,1,\cdots) \tag{2.30}$$

其中 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$ 任选。由迭代公式(2.30)所表示的迭代法称为 Gauss - Seidel(高斯-赛德尔)迭代法，简称 GS 法，它的迭代矩阵是

$$\boldsymbol{G}_G = -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{U}$$

实际使用 GS 法时，为避免求 $\boldsymbol{D}+\boldsymbol{L}$ 的逆矩阵，把迭代公式(2.30)改写为

$$\begin{aligned} (\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}^{(k+1)} &= -\boldsymbol{U}\boldsymbol{x}^{(k)} + \boldsymbol{b} \\ \boldsymbol{D}\boldsymbol{x}^{(k+1)} &= -\boldsymbol{L}\boldsymbol{x}^{(k+1)} - \boldsymbol{U}\boldsymbol{x}^{(k)} + \boldsymbol{b} \\ \boldsymbol{x}^{(k+1)} &= -\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{x}^{(k+1)} - \boldsymbol{D}^{-1}\boldsymbol{U}\boldsymbol{x}^{(k)} + \boldsymbol{D}^{-1}\boldsymbol{b} \end{aligned}$$

其分量形式为

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{(k+1)} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{(k)} + \frac{b_i}{a_{ii}} \quad (i=1,2,\cdots,n;k=0,1,\cdots) \tag{2.31}$$

由定理 2.8 得到

定理 2.13　GS 法收敛的充分必要条件是 $\rho(\boldsymbol{G}_G)<1$。

由定理 2.9 得到

定理 2.14　如果 $\|\boldsymbol{G}_G\|<1$，则 GS 法收敛。

下面两个定理都是直接用原方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 判断收敛。

定理 2.15　如果方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 是主对角线按行(或按列)严格占优阵，则用 GS 法求解必收敛。

此定理在这里不证明了，因为它是后面定理 2.20 的一个直接结果。

定理 2.16　如果方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 是正定矩阵，则用 GS 法求解必收敛。

此定理在这里也不证明，因为它是后面定理 2.21 的一个直接结果。

例 9　证明用 GS 法求解例 6 的方程组必收敛，并求解，要求 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_\infty\leqslant 10^{-5}$。

解　迭代矩阵为

$$\boldsymbol{G}_G = -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{U} = -\frac{1}{300}\begin{bmatrix} 0 & -60 & -60 \\ 0 & -12 & -42 \\ 0 & -28 & -48 \end{bmatrix}$$

因为 $\|\boldsymbol{G}_G\|_1=\dfrac{150}{300}<1$，所以，GS 法必收敛。

由式(2.31)，迭代公式为

$$\begin{cases} x_1^{(k+1)} = 0.2x_2^{(k)} + 0.2x_3^{(k)} + 0.1 \\ x_2^{(k+1)} = 0.2x_1^{(k+1)} + 0.1x_3^{(k)} + 0.05 \\ x_3^{(k+1)} = \dfrac{1}{3}x_1^{(k+1)} + \dfrac{2}{3}x_2^{(k+1)} + \dfrac{1}{3} \\ \qquad [(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^{\mathrm{T}} \text{任选}; k = 0, 1, \cdots] \end{cases}$$

计算结果见表 2-2。

表 2-2 例 9 计算结果

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.0	0.0	0.0
1	0.100 000	0.070 000	0.413 333
2	0.196 667	0.130 667	0.486 000
3	0.223 333	0.143 267	0.503 289
4	0.229 311	0.146 191	0.507 231
5	0.230 684	0.146 860	0.508 134
6	0.230 999	0.147 013	0.508 341
7	0.231 071	0.147 048	0.508 389
8	0.231 087	0.147 056	0.508 399
9	0.231 091	0.147 058	0.508 402

因 $\|\boldsymbol{x}^{(9)} - \boldsymbol{x}^{(8)}\|_\infty = 0.4 \times 10^{-5}$，故得方程组的解为

$$x_1^* \approx 0.231\,091, \quad x_2^* \approx 0.147\,058, \quad x_3^* \approx 0.508\,402$$

例 10 证明用 GS 法求解例 7 的方程组必收敛。

证 迭代矩阵为

$$\boldsymbol{G}_{\mathrm{G}} = -\begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & -2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & -0.5 \\ 0 & -0.5 & -0.5 \\ 0 & 0 & -0.5 \end{bmatrix}$$

可见，$\boldsymbol{G}_{\mathrm{G}}$ 的特征值为 $\lambda_1 = 0, \lambda_2 = \lambda_3 = -0.5$，即 $\rho(\boldsymbol{G}_{\mathrm{G}}) = 0.5 < 1$，所以，GS 法必收敛。

证毕。

用某种迭代法求解线性方程组是否收敛只取决于方程组的系数矩阵 $\boldsymbol{A}$。前面已看到，对于一个给定的系数矩阵 $\boldsymbol{A}$，Jacobi 迭代法和 GS 法可能都收敛，也可能都不收敛，还可能 Jacobi 迭代法收敛而 GS 法不收敛，或者 GS 法收敛而 Jacobi 迭代法不收敛。在两者都收敛的情况下，可能 Jacobi 迭代法收敛得快一些，也可能 GS 法收敛得快一些。有时候，对系数矩阵 $\boldsymbol{A}$ 作适当的改变就能变不收敛为收敛。

例如对于方程组

$$\begin{cases} 2x_1 - 5x_2 = 1 \\ 10x_1 - 4x_2 = 3 \end{cases}$$

Jacobi 迭代法的迭代矩阵是

$$\boldsymbol{G}_{\mathrm{J}} = \begin{bmatrix} 0 & 2.5 \\ 2.5 & 0 \end{bmatrix}$$

它的两个特征值为 $\lambda_{1,2}=\pm 2.5$。由于 $\rho(\boldsymbol{G}_J)>1$，所以 Jacobi 迭代法不收敛。GS 法的迭代矩阵是

$$\boldsymbol{G}_G=-\begin{bmatrix}2 & 0\\10 & -4\end{bmatrix}^{-1}\begin{bmatrix}0 & -5\\0 & 0\end{bmatrix}=\begin{bmatrix}0 & 2.5\\0 & 6.25\end{bmatrix}$$

可见，$\rho(\boldsymbol{G}_G)=6.25>1$，故 GS 法也不收敛。如果把原方程组中的两个方程交换次序，则得

$$\begin{cases}10x_1-4x_2=3\\2x_1-5x_2=1\end{cases}$$

由于系数矩阵是主对角线按行严格占优阵，所以，用 Jacobi 迭代法和 GS 法求解，都是收敛的。

在计算机上实际计算时，GS 法只需要一套存放迭代向量 $\boldsymbol{x}^{(k)}$ 的单元，而 Jacobi 迭代法却需要两套。

2.4.4 逐次超松弛迭代法

设方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 满足 $a_{ii}\neq 0(i=1,2,\cdots,n)$。把 $\boldsymbol{A}$ 分解为

$$\boldsymbol{A}=\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}+\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}$$

其中实常数 $\omega>0$ 称为松弛因子。在迭代法一般形式(2.24)中，取

$$\boldsymbol{N}=\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L},\quad \boldsymbol{P}=-\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]$$

形成以下的迭代公式

$$\begin{aligned}\boldsymbol{x}^{(k+1)}=&-\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]\boldsymbol{x}^{(k)}+\\&\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\boldsymbol{b}\qquad(k=0,1,\cdots)\end{aligned}\tag{2.32}$$

其中 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$ 任选。由迭代公式(2.32)所表示的迭代法称为逐次超松弛迭代法(Successive Over Relaxation Method)，简称 SOR 方法。当 $\omega=1$ 时，式(2.32)就是 GS 法。SOR 方法(2.32)的迭代矩阵是

$$\boldsymbol{G}_S=-\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]$$

在实际使用 SOR 方法时，为避免求 $\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}$ 的逆矩阵，把其迭代公式(2.32)作如下的变动：用 $\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}$ 左乘式(2.32)两端，并经整理，得

$$\begin{aligned}\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)\boldsymbol{x}^{(k+1)}&=-\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]\boldsymbol{x}^{(k)}+\boldsymbol{b}\\\frac{1}{\omega}\boldsymbol{D}\boldsymbol{x}^{(k+1)}&=-\boldsymbol{L}\boldsymbol{x}^{(k+1)}-\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]\boldsymbol{x}^{(k)}+\boldsymbol{b}\\\boldsymbol{x}^{(k+1)}&=\omega\left\{-\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{x}^{(k+1)}-\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{U}\right]\boldsymbol{x}^{(k)}+\boldsymbol{D}^{-1}\boldsymbol{b}\right\}\end{aligned}$$

它的分量形式是

$$x_i^{(k+1)}=\omega\left[-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{(k+1)}-\left(1-\frac{1}{\omega}\right)x_i^{(k)}-\sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{(k)}+\frac{b_i}{a_{ii}}\right]$$

$$(i=1,2,\cdots,n;\ k=0,1,\cdots)\tag{2.33}$$

式(2.33)是 SOR 方法实际计算公式。

由定理 2.8 得到

定理 2.17 SOR 方法收敛的充分必要条件是 $\rho(\boldsymbol{G}_S)<1$。

由定理 2.9 得到

定理 2.18 如果 $\|\boldsymbol{G}_S\|<1$,则 SOR 方法收敛。

定理 2.19 SOR 方法收敛的必要条件是 $0<\omega<2$。

证 因为

$$\det\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}=\omega^n\prod_{i=1}^{n}\frac{1}{a_{ii}}$$

$$\det\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]=\left(1-\frac{1}{\omega}\right)^n\prod_{i=1}^{n}a_{ii}$$

所以

$$\det\boldsymbol{G}_S=(-1)^n\cdot\det\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\cdot\det\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]=(1-\omega)^n$$

因而 $\boldsymbol{G}_S$ 的所有特征值 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 满足

$$\lambda_1\lambda_2\cdots\lambda_n=(1-\omega)^n$$

如果 SOR 方法收敛,则 $\rho(\boldsymbol{G}_S)<1$,从而有

$$|1-\omega|<1,\quad 0<\omega<2$$

证毕。

定理 2.20 如果方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 是主对角线按行(或按列)严格占优阵,则用 $0<\omega\leqslant1$的 SOR 方法求解必收敛。

证 由定理条件知$\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}$存在。假定 $0<\omega\leqslant1$ 的 SOR 方法不收敛,则根据定理 2.8,$\boldsymbol{G}_S$ 必有一个特征值 μ,$|\mu|\geqslant1$,并且应有

$$\det(\mu\boldsymbol{I}-\boldsymbol{G}_S)=0$$

注意到

$$\begin{aligned}\mu\boldsymbol{I}-\boldsymbol{G}_S=&\mu\boldsymbol{I}+\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]=\\&\mu\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\left\{\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)+\frac{1}{\mu}\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{U}\right]\right\}=\\&\mu\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\left\{\left[\frac{1}{\mu}\left(1-\frac{1}{\omega}\right)+\frac{1}{\omega}\right]\boldsymbol{D}+\boldsymbol{L}+\frac{1}{\mu}\boldsymbol{U}\right\}\end{aligned}$$

记

$$\boldsymbol{B}=\left[\frac{1}{\mu}\left(1-\frac{1}{\omega}\right)+\frac{1}{\omega}\right]\boldsymbol{D}+\boldsymbol{L}+\frac{1}{\mu}\boldsymbol{U}$$

则

$$\det(\mu\boldsymbol{I}-\boldsymbol{G}_S)=\mu^n\cdot\det\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\cdot\det\boldsymbol{B}$$

显然

$$\mu^n\neq0,\quad\det\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\neq0$$

设 $\boldsymbol{B}=[b_{ij}]_{n\times n}$，则

$$b_{ii}=\left[\frac{1}{\mu}\left(1-\frac{1}{\omega}\right)+\frac{1}{\omega}\right]a_{ii}$$

$$b_{ij}=\begin{cases}a_{ij}, & i>j\\ \dfrac{1}{\mu}a_{ij}, & i<j\end{cases}$$

因 $|\mu|\geqslant 1, 0<\omega\leqslant 1$，故

$$\left|\frac{1}{\mu}\left(1-\frac{1}{\omega}\right)+\frac{1}{\omega}\right|=\left|1+\left(1-\frac{1}{\mu}\right)\left(\frac{1}{\omega}-1\right)\right|\geqslant 1$$

$$|b_{ii}|\geqslant|a_{ii}|$$

$$|a_{ij}|\geqslant|b_{ij}|,\quad i\neq j$$

于是，由 $\boldsymbol{A}$ 是主对角线按行（或按列）严格占优阵可推知 $\boldsymbol{B}$ 也是主对角线按行（或按列）严格占优阵，因而 $\det\boldsymbol{B}\neq 0$，并得出

$$\det(\mu\boldsymbol{I}-\boldsymbol{G}_{\mathrm{S}})\neq 0$$

所出现的矛盾证实定理的结论成立。

证毕。

当 $\omega=1$ 时，由此定理可得定理 2.15。

定理 2.21　如果方程组(2.2)的系数矩阵 $\boldsymbol{A}$ 是正定矩阵，则用 $0<\omega<2$ 的 SOR 方法求解必收敛。

证　因 $\boldsymbol{A}$ 是正定矩阵，故 $a_{ii}>0(i=1,2,\cdots,n)$，$\boldsymbol{D}$ 也是正定矩阵，$\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}$ 存在；又因 $\boldsymbol{A}$ 对称，故 $\boldsymbol{U}=\boldsymbol{L}^{\mathrm{T}}$，迭代矩阵为

$$\boldsymbol{G}_{\mathrm{S}}=-\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)^{-1}\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{L}^{\mathrm{T}}\right]$$

设 λ 是 $\boldsymbol{G}_{\mathrm{S}}$ 的任一特征值，相应的特征向量为 $\boldsymbol{y}$，则有

$$\boldsymbol{G}_{\mathrm{S}}\boldsymbol{y}=\lambda\boldsymbol{y}$$

$$-\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{L}^{\mathrm{T}}\right]\boldsymbol{y}=\lambda\left(\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}\right)\boldsymbol{y}\tag{2.34}$$

把

$$-\left[\left(1-\frac{1}{\omega}\right)\boldsymbol{D}+\boldsymbol{L}^{\mathrm{T}}\right]=\frac{1}{2}\left[\left(\frac{2}{\omega}-1\right)\boldsymbol{D}-\boldsymbol{A}+\boldsymbol{L}-\boldsymbol{L}^{\mathrm{T}}\right]$$

$$\frac{1}{\omega}\boldsymbol{D}+\boldsymbol{L}=\frac{1}{2}\left[\left(\frac{2}{\omega}-1\right)\boldsymbol{D}+\boldsymbol{A}+\boldsymbol{L}-\boldsymbol{L}^{\mathrm{T}}\right]$$

代入式(2.34)，并且用 $\boldsymbol{y}^{\mathrm{H}}$ 左乘所得等式两边，得到

$$\boldsymbol{y}^{\mathrm{H}}\left[\left(\frac{2}{\omega}-1\right)\boldsymbol{D}-\boldsymbol{A}+\boldsymbol{L}-\boldsymbol{L}^{\mathrm{T}}\right]\boldsymbol{y}=\lambda\boldsymbol{y}^{\mathrm{H}}\left[\left(\frac{2}{\omega}-1\right)\boldsymbol{D}+\boldsymbol{A}+\boldsymbol{L}-\boldsymbol{L}^{\mathrm{T}}\right]\boldsymbol{y}\tag{2.35}$$

因 $\boldsymbol{D}$ 和 $\boldsymbol{A}$ 都是正定矩阵，$\boldsymbol{y}\neq\boldsymbol{0}$，所以

$$\boldsymbol{y}^{\mathrm{H}}\boldsymbol{D}\boldsymbol{y}=d>0,\quad \boldsymbol{y}^{\mathrm{H}}\boldsymbol{A}\boldsymbol{y}=a>0$$

又设 $\boldsymbol{y}^{\mathrm{H}}\boldsymbol{L}\boldsymbol{y}=\alpha+\mathrm{i}\beta$，则 $\boldsymbol{y}^{\mathrm{H}}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{y}=\alpha-\mathrm{i}\beta$。于是，由等式(2.35)得到

$$\lambda=\frac{\left(\frac{2}{\omega}-1\right)d-a+2\beta\mathrm{i}}{\left(\frac{2}{\omega}-1\right)d+a+2\beta\mathrm{i}}$$

因 $0<\omega<2$，故 $\frac{2}{\omega}-1>0$，因此有

$$\left|\left(\frac{2}{\omega}-1\right)d-a\right|<\left(\frac{2}{\omega}-1\right)d+a$$

由此可知 $|\lambda|<1$，从而 $\rho(\boldsymbol{G}_S)<1$，SOR 方法收敛。

证毕。

当 $\omega=1$ 时，由此定理可得定理 2.16。

例 11 试用 $\omega=1.25$ 的 SOR 方法求解方程组

$$\begin{cases}4x_1+3x_2 & =16\\ 3x_1+4x_2-x_3 & =20\\ -x_2+4x_3 & =-12\end{cases}$$

要求 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_\infty<0.000\,05$。

解 此方程组的系数矩阵

$$\boldsymbol{A}=\begin{bmatrix}4 & 3 & 0\\ 3 & 4 & -1\\ 0 & -1 & 4\end{bmatrix}$$

是正定矩阵，故用 $\omega=1.25$ 的 SOR 方法求解必收敛。迭代公式为

$$\begin{cases}x_1^{(k+1)}=1.25(-0.2x_1^{(k)}-0.75x_2^{(k)}+4)\\ x_2^{(k+1)}=1.25(-0.75x_1^{(k+1)}-0.2x_2^{(k)}+0.25x_3^{(k)}+5)\\ x_3^{(k+1)}=1.25(0.25x_2^{(k+1)}-0.2x_3^{(k)}-3)\\ \qquad (k=0,1,\cdots)\end{cases}$$

计算结果见表 2-3。由于 $\|\boldsymbol{x}^{(12)}-\boldsymbol{x}^{(11)}\|_\infty=0.000\,04<0.000\,05$，故得方程组的解为

$$x_1^*\approx 1.500\,01,\quad x_2^*\approx 3.333\,33,\quad x_3^*\approx -2.166\,67$$

表 2-3 例 11 计算结果

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.0	0.0	0.0
1	5.000 00	1.562 50	−3.261 72
2	2.285 16	2.697 75	−2.091 52
3	1.899 57	2.779 63	−2.358 49
4	1.919 20	3.018 81	−2.217 00
5	1.690 07	3.218 04	−2.190 11
6	1.560 57	3.298 05	−2.171 83
7	1.517 94	3.323 72	−2.168 38
8	1.504 53	3.330 95	−2.166 98
9	1.501 10	3.332 80	−2.166 76
10	1.500 23	3.333 22	−2.166 68
11	1.500 05	3.333 31	−2.166 67
12	1.500 01	3.333 33	−2.166 67

使用 SOR 方法求解线性方程组，首先要选择松弛因子 ω。所选出的 ω，不仅要使迭代法收

敛，而且应有高的收敛速度。在例 11 中，如果选 $\omega=1$，也就是用 GS 法，那么要得到同样精度的结果需要迭代 22 次。关于如何选取最佳松弛因子的问题，至今仍没有有效地解决。在实际计算工作中，经常采用试算的方法寻找较好的松弛因子。特别在同时求解多个具有相同系数矩阵的线性方程组时，通过试验找出恰当的松弛因子能大大提高计算效率。

习　题

1. 在三位十进制的限制下，试分别用顺序 Gauss 消去法和列主元素 Gauss 消去法求解方程组

$$\begin{cases} 0.5x_1 + 1.1\,x_2 + \ 3.1x_3 = 6 \\ 5x_1 + 0.96x_2 + \ 6.5x_3 = 0.96 \\ 2x_1 + \ 4.5x_2 + 0.36x_3 = 0.02 \end{cases}$$

（精确解为 $x_1=-2.6, x_2=1, x_3=2$）

2. 顺序 Gauss 消去法中，$k=1$ 的消元过程相当于用一个 $n\times n$ 非奇异矩阵 $\boldsymbol{P}_1$ 左乘增广矩阵 $[\boldsymbol{A},\boldsymbol{b}]$，即

$$\boldsymbol{P}_1[\boldsymbol{A},\boldsymbol{b}] = [\boldsymbol{A}^{(2)},\boldsymbol{b}^{(2)}]$$

试问：$\boldsymbol{P}_1$ 是什么样的矩阵？

3. 试问：交换增广矩阵 $[\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)}]$ 的第 k 行与第 i_k 行，相当于用一个什么样的矩阵 $\boldsymbol{Q}_k$ 左乘 $[\boldsymbol{A}^{(k)},\boldsymbol{b}^{(k)}]$？

4. 试统计 Doolittle 分解法求解 n 元线性方程组所需的乘除法运算次数。

5. 试用不选主元的 Doolittle 分解法求解方程组

$$\begin{cases} 3.2x_1 - 1.4x_2 + \ x_3 = 2.5 \\ 6.4x_1 + 2.8x_2 + 3x_3 = 1.8 \\ 5.5x_1 + 5.2x_2 - 4x_3 = 7.2 \end{cases}$$

6. 试用选主元的 Doolittle 分解法求解方程组

$$\begin{bmatrix} 1 & 8 & 2 & 3 \\ -6 & -3 & 8 & 1 \\ 2 & 4 & 4 & 2 \\ 10 & 5 & -5 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 40 \\ -50 \\ 80 \end{bmatrix}$$

7. 设上下半带宽分别为 s 和 r 的 $n\times n$ 带状矩阵 $\boldsymbol{A}$ 的带内元素已存放在 $(r+s+1)\times n$ 的矩阵 $\boldsymbol{C}$ 中，已知 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，$\boldsymbol{u}=\boldsymbol{A}\boldsymbol{y}$，试写出由矩阵 $\boldsymbol{C}$ 的元素计算 $\boldsymbol{u}$ 的分量 $u_i(i=1,2,\cdots,n)$ 的计算公式。

8. 试用追赶法求解下列方程组

$$\begin{bmatrix} 8 & -1 & & & \\ 2 & 8 & 1 & & \\ & 1 & 8 & -1 & \\ & & 2 & 8 & 1 \\ & & & 1 & 8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2.2 \\ -2.54 \\ 8.26 \\ 8.32 \\ -4.3 \end{bmatrix}$$

(精确解为 $x_1=0.21, x_2=-0.52, x_3=1.2, x_4=0.82, x_5=-0.64$)

9. 在八位十进制的限制下,试用列主元素 Gauss 消去法求解方程组

$$\begin{cases}0.216\,1x_1+0.144\,1x_2=0.144\\1.296\,9x_1+0.864\,8x_2=0.864\,2\end{cases}$$

(精确解为 $x_1=2, x_2=-2$),并求它的系数矩阵的(行范数)条件数,计算结果说明什么问题?

10. 设 $\boldsymbol{A}$ 是正交矩阵,试证:$\boldsymbol{A}$ 的(谱范数)条件数等于 1。

11. 设 $\boldsymbol{A}$ 是非奇异矩阵,$\boldsymbol{b}\neq\boldsymbol{0}$,$\tilde{\boldsymbol{x}}$ 是方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的一个近似解,用 $\boldsymbol{x}^*$ 表示其精确解,记 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$(称为残差),证明:

$$\frac{\|\boldsymbol{x}^*-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}^*\|}\leqslant \mathrm{cond}(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}$$

12. 设

$$\boldsymbol{A}=\begin{bmatrix}1 & 0.99\\0.99 & 0.98\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}1\\1\end{bmatrix}$$

已知方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的精确解为 $\boldsymbol{x}^*=(100,-100)^{\mathrm{T}}$,

(1) 计算 $\boldsymbol{A}$ 的(行范数)条件数;

(2) 取 $\tilde{\boldsymbol{x}}=(1,0)^{\mathrm{T}}$,计算残差 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$;

(3) 取 $\tilde{\boldsymbol{x}}=(100.5,-99.5)^{\mathrm{T}}$,计算残差 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}$。

本题计算结果说明什么问题?

13. 试判断下列迭代公式产生的向量序列 $\{\boldsymbol{x}^{(k)}\}$ 是否收敛,其中 d_1,d_2,d_3 是常数,$\boldsymbol{x}^{(0)}\in\mathbf{R}^3$ 任取。

(1) $$\begin{bmatrix}x_1^{(k+1)}\\x_2^{(k+1)}\\x_3^{(k+1)}\end{bmatrix}=\begin{bmatrix}1&2&2\\2&1&2\\2&2&1\end{bmatrix}\begin{bmatrix}x_1^{(k)}\\x_2^{(k)}\\x_3^{(k)}\end{bmatrix}+\begin{bmatrix}d_1\\d_2\\d_3\end{bmatrix}\quad(k=0,1,2,\cdots);$$

(2) $$\begin{bmatrix}x_1^{(k+1)}\\x_2^{(k+1)}\\x_3^{(k+1)}\end{bmatrix}=\begin{bmatrix}0.2&0.5&-0.1\\0&-0.8&0.1\\-0.4&0.2&0.3\end{bmatrix}\begin{bmatrix}x_1^{(k)}\\x_2^{(k)}\\x_3^{(k)}\end{bmatrix}+\begin{bmatrix}d_1\\d_2\\d_3\end{bmatrix}\quad(k=0,1,2,\cdots)。$$

14. 试证明下列迭代公式产生的向量序列 $\{\boldsymbol{x}^{(k)}\}$ 必收敛;并问:此迭代公式收敛于哪个线性方程组的解? 其中 $\boldsymbol{x}^{(0)}\in\mathbf{R}^3$ 任取。

$$\begin{cases}x_1^{(k+1)}=0.4x_2^{(k)}+0.2x_3^{(k)}+1\\x_2^{(k+1)}=0.25x_1^{(k+1)}-x_2^{(k)}-0.5x_3^{(k)}+2\\x_3^{(k+1)}=0.25x_1^{(k+1)}+0.5x_2^{(k)}+0.8x_3^{(k)}+3\\\qquad(k=0,1,2,\cdots)\end{cases}$$

15. 设有方程组

$$\begin{cases}5x_1+2x_2+x_3=-12\\-x_1+4x_2+2x_3=10\\2x_1-5x_2+10x_3=1\end{cases}$$

用 Jacobi 迭代法求解,要求满足 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_\infty\leqslant 10^{-4}$ 时迭代终止,并判断迭代过程是否收敛?

16. 讨论用 Jacobi 迭代法求解方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 的收敛性,其中

(1) $A=\begin{bmatrix} 2 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -2 \end{bmatrix}$；　(2) $A=\begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}$；

(3) $A=\begin{bmatrix} 1 & -2 & 2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{bmatrix}$；　(4) $A=\begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 1 & -1 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$。

17. 用 Gauss - Seidel 迭代法求解第 15 题的方程组，当满足 $\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \|_\infty < 10^{-4}$ 时结束迭代，并判断迭代过程是否收敛？

18. 讨论用 Gauss - Seidel 迭代法求解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的收敛性，其中 $\boldsymbol{A}$ 由第 16 题给出。

19. 为求解方程组

$$\begin{cases} x_1 + 2x_2 - 5x_3 = 10 \\ 10x_1 - 2x_2 = 3 \\ 2x_1 + 10x_2 - x_3 = 15 \end{cases}$$

试写出一个必收敛的迭代公式，并说明收敛的理由。

20. 设 $\boldsymbol{A}$ 是 2×2 矩阵，且 $a_{11}\neq 0, a_{22}\neq 0$，试证：对方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，Jacobi 迭代法和 Gauss - Seidel 迭代法同时收敛和发散。

21. 设

$$\boldsymbol{A} = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

其中 a 为实数，

(1) a 取何值时，用 Jacobi 迭代法求解 $\boldsymbol{Ax}=\boldsymbol{b}$ 收敛？

(2) a 取何值时，用 Gauss - Seidel 迭代法求解 $\boldsymbol{Ax}=\boldsymbol{b}$ 收敛？

22. 试由系数矩阵 $\boldsymbol{A}$ 直接判定 Gauss - Seidel 迭代法求解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 必收敛，其中

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & -2 \\ 0 & -2 & 5 \end{bmatrix}$$

23. 试用 SOR 迭代法（取 $\omega=0.9$）求解方程组

$$\begin{cases} 5x_1 + 2x_2 + x_3 = -12 \\ -x_1 + 4x_2 + 2x_3 = 20 \\ 2x_1 - 3x_2 + 10x_3 = 3 \end{cases}$$

当满足 $\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \|_\infty < 10^{-5}$ 时迭代终止，并证明迭代过程是收敛的。

24. 试用 SOR 迭代法（取 $\omega=1.25$）求解方程组

$$\begin{cases} 2x_1 + 2x_2 - 2x_3 = 1 \\ 2x_1 + 5x_2 - 4x_3 = 2 \\ -2x_1 - 4x_2 + 5x_3 = 0 \end{cases}$$

当满足 $\| \boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)} \|_\infty < 10^{-5}$ 时迭代终止，并证明迭代过程是收敛的。

第3章　矩阵特征值与特征向量的计算

设 $\boldsymbol{A}$ 是 $n\times n$ 矩阵，如果数 λ 和 n 维非零向量 $\boldsymbol{x}$ 满足

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$$

则称 λ 为矩阵 $\boldsymbol{A}$ 的一个特征值，$\boldsymbol{x}$ 称为矩阵 $\boldsymbol{A}$ 的属于 λ 的特征向量。

在工程技术中，计算矩阵的特征值和特征向量主要使用数值解法。本章将阐述其中几种，并且只限于讨论实矩阵的情形。

3.1　幂法和反幂法

3.1.1　幂　法

幂法主要用于计算矩阵的按模为最大的特征值和相应的特征向量。

设 $n\times n$ 实矩阵 $\boldsymbol{A}$ 具有 n 个线性无关的特征向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$，其相应的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 满足不等式

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant |\lambda_n| \tag{3.1}$$

其中 $\boldsymbol{A}\boldsymbol{x}_i=\lambda_i\boldsymbol{x}_i(i=1,2,\cdots,n)$。现在要求出 λ_1 和相应的特征向量。

任取一 n 维非零向量 $\boldsymbol{u}_0$，从 $\boldsymbol{u}_0$ 出发，按照如下的递推公式

$$\boldsymbol{u}_k = \boldsymbol{A}\boldsymbol{u}_{k-1} \quad (k = 1,2,\cdots) \tag{3.2}$$

可产生一个向量序列 $\{\boldsymbol{u}_k\}$，分析这一序列的收敛情况，可从中找出计算 λ_1 和相应的特征向量的方法。

因 n 维向量组 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 线性无关，故对于向量 $\boldsymbol{u}_0$，必存在唯一的不全为零的数组 $\alpha_1,\alpha_2,\cdots,\alpha_n$，使得

$$\boldsymbol{u}_0 = \alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_n\boldsymbol{x}_n$$

由式(3.2)可得

$$\begin{aligned}\boldsymbol{u}_k &= \boldsymbol{A}\boldsymbol{u}_{k-1} = \boldsymbol{A}^2\boldsymbol{u}_{k-2} = \cdots = \boldsymbol{A}^k\boldsymbol{u}_0 = \\ &\alpha_1\boldsymbol{A}^k\boldsymbol{x}_1 + \alpha_2\boldsymbol{A}^k\boldsymbol{x}_2 + \cdots + \alpha_n\boldsymbol{A}^k\boldsymbol{x}_n = \\ &\alpha_1\lambda_1^k\boldsymbol{x}_1 + \alpha_2\lambda_2^k\boldsymbol{x}_2 + \cdots + \alpha_n\lambda_n^k\boldsymbol{x}_n = \\ &\lambda_1^k\left[\alpha_1\boldsymbol{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{x}_2 + \cdots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{x}_n\right]\end{aligned} \tag{3.3}$$

设 $\alpha_1\neq0$。由于有式(3.1)成立，故从式(3.3)可看出，当 k 充分大时，有

$$\boldsymbol{u}_k \approx \lambda_1^k\alpha_1\boldsymbol{x}_1$$

因为矩阵的特征向量与任何一个非零常数相乘之后仍是该矩阵的属于同一个特征值的特征向量，所以，当 k 充分大时，由迭代公式(3.2)产生的 $\boldsymbol{u}_k$ 可近似地作为矩阵 $\boldsymbol{A}$ 的属于 λ_1 的特征向量。迭代公式(3.2)实质上是

$$\boldsymbol{u}_k = \boldsymbol{A}^k\boldsymbol{u}_0$$

故称这种迭代法为幂法。如果所选取的 $\boldsymbol{u}_0$ 使得 $\alpha_1=0$，那么由于计算过程有舍入误差的影响，必然会在迭代的某一步产生这样的 $\tilde{\boldsymbol{u}}_k$，它在 $\boldsymbol{x}_1$ 方向上的分量不为零。这时，相当于以 $\tilde{\boldsymbol{u}}_k$ 为初始向量重新开始迭代。

实际计算时，为了避免迭代向量 $\boldsymbol{u}_k$ 的模过大（当 $|\lambda_1|>1$）或过小（当 $|\lambda_1|<1$），通常每迭代一次都对 $\boldsymbol{u}_k$ 进行归一化，使其范数（$\|\cdot\|_2$ 或 $\|\cdot\|_\infty$）等于 1。因此，实际使用的迭代公式是

$$\begin{cases}\boldsymbol{y}_{k-1}=\dfrac{\boldsymbol{u}_{k-1}}{\|\boldsymbol{u}_{k-1}\|}\\ \boldsymbol{u}_k=\boldsymbol{A}\boldsymbol{y}_{k-1}\end{cases}\quad(k=1,2,\cdots)\tag{3.4}$$

由迭代公式(3.4)可知

$$\boldsymbol{u}_k=\frac{\boldsymbol{A}\boldsymbol{u}_{k-1}}{\|\boldsymbol{u}_{k-1}\|}=\frac{\boldsymbol{A}^2\boldsymbol{u}_{k-2}}{\|\boldsymbol{A}\boldsymbol{u}_{k-2}\|}=\cdots=\frac{\boldsymbol{A}^k\boldsymbol{u}_0}{\|\boldsymbol{A}^{k-1}\boldsymbol{u}_0\|}$$

因而

$$\boldsymbol{y}_k=\frac{\boldsymbol{u}_k}{\|\boldsymbol{u}_k\|}=\frac{\boldsymbol{A}^k\boldsymbol{u}_0}{\|\boldsymbol{A}^k\boldsymbol{u}_0\|}=$$

$$\left(\frac{\lambda_1}{|\lambda_1|}\right)^k\frac{\alpha_1\boldsymbol{x}_1+\alpha_2\left(\dfrac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{x}_2+\cdots+\alpha_n\left(\dfrac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{x}_n}{\left\|\alpha_1\boldsymbol{x}_1+\alpha_2\left(\dfrac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{x}_2+\cdots+\alpha_n\left(\dfrac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{x}_n\right\|}\tag{3.5}$$

由于式(3.1)成立，故当 $k\to\infty$ 时，若 $\lambda_1>0$，就有 $\boldsymbol{y}_k\to\dfrac{\alpha_1\boldsymbol{x}_1}{\|\alpha_1\boldsymbol{x}_1\|}$，即 $\boldsymbol{y}_k$ 有确定的极限；若 $\lambda_1<0$，则 $\boldsymbol{y}_k$ 的各个分量之模有确定的极限，而 $\boldsymbol{y}_k$ 各个分量的正负号每迭代一次就改变一次。无论是哪一种情况都说明，当 k 充分大时，由式(3.4)得到的 $\boldsymbol{y}_{k-1}$ 可以近似地作为 $\boldsymbol{A}$ 的属于 λ_1 的特征向量。

关于 λ_1 的计算，有两种方法可供选择。第一种方法，在式(3.4)中使用范数 $\|\cdot\|_2$，并且令

$$\beta_k=\boldsymbol{y}_{k-1}^{\mathrm{T}}\boldsymbol{u}_k\tag{3.6}$$

那么，由于 $\boldsymbol{u}_k=\boldsymbol{A}\boldsymbol{y}_{k-1}$，并根据式(3.5)可得

$$\lim_{k\to\infty}\beta_k=\frac{\alpha_1\boldsymbol{x}_1^{\mathrm{T}}}{\|\alpha_1\boldsymbol{x}_1\|_2}\boldsymbol{A}\frac{\alpha_1\boldsymbol{x}_1}{\|\alpha_1\boldsymbol{x}_1\|_2}=\frac{\alpha_1^2\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1}{\|\alpha_1\boldsymbol{x}_1\|_2^2}\lambda_1=\lambda_1$$

把式(3.4)和(3.6)结合在一起，得到第一种幂法迭代格式：

$$\begin{cases}\text{任取非零向量 }\boldsymbol{u}_0\in\mathbf{R}^n\\ \eta_{k-1}=\sqrt{\boldsymbol{u}_{(k-1)}^{\mathrm{T}}\boldsymbol{u}_{k-1}}\\ \boldsymbol{y}_{k-1}=\boldsymbol{u}_{k-1}/\eta_{k-1}\\ \boldsymbol{u}_k=\boldsymbol{A}\boldsymbol{y}_{k-1}\\ \beta_k=\boldsymbol{y}_{k-1}^{\mathrm{T}}\boldsymbol{u}_k\\ \qquad(k=1,2,\cdots)\end{cases}\tag{3.7}$$

当 $|\beta_k-\beta_{k-1}|/|\beta_k|\leqslant\varepsilon$（允许误差）时，迭代终止，以当前的 β_k 作为 λ_1 的近似值，以 $\boldsymbol{y}_{k-1}$ 作为 $\boldsymbol{A}$ 的属于 λ_1 的特征向量。

第二种方法是在式(3.4)中使用范数 $\|\cdot\|_\infty$，并且令

$$\beta_k = \frac{e_r^{\mathrm{T}} u_k}{e_r^{\mathrm{T}} y_{k-1}} \tag{3.8}$$

这里假定 u_{k-1} 的第 r 个分量为模最大的分量，当 k 足够大之后，r 保持定值；e_r 是 n 维基本单位向量，它的第 r 个分量为 1，其余分量为零。由于 $u_k = Ay_{k-1}$，并根据式(3.5)可得

$$\lim_{k\to\infty} \beta_k = \frac{e_r^{\mathrm{T}} A x_1}{e_r^{\mathrm{T}} x_1} = \lambda_1$$

把式(3.4)和(3.8)结合在一起，得到第二种幂法迭代格式：

$$\begin{cases} \text{任取非零向量 } u_0 = (h_1^{(0)}, \cdots, h_n^{(0)})^{\mathrm{T}} \\ |h_r^{(k-1)}| = \max\limits_{1\leqslant j\leqslant n} |h_j^{(k-1)}| \\ y_{k-1} = u_{k-1} / |h_r^{(k-1)}| \\ u_k = Ay_{k-1} = (h_1^{(k)}, \cdots, h_n^{(k)})^{\mathrm{T}} \\ \beta_k = \mathrm{sgn}(h_r^{k-1}) h_r^{(k)} \\ \qquad (k = 1, 2, \cdots) \end{cases} \tag{3.9}$$

终止迭代的控制也用 $|\beta_k - \beta_k| / |\beta_k| \leqslant \varepsilon$，当前的 β_k 和 y_{k-1} 即分别作为 λ_1 和与其相应的特征向量。在迭代格式(3.9)中，$|h_r^{k-1}| = \| u_{k-1} \|_\infty$，$\mathrm{sgn}(h_r^{(k-1)}) = e_r^{\mathrm{T}} y_{k-1}$，$h_r^{(k)} = e_r^{\mathrm{T}} u_k$。

两种迭代格式(3.7)和(3.9)相比较，格式(3.7)编制程序容易，迭代一次所需时间较短；格式(3.9)每迭代一次都要判断 u_{k-1} 的第几个分量的模最大，因而所需时间较长，但它在计算过程中舍入误差的影响比格式(3.7)小。

设矩阵 A 的特征值 $\lambda_i (i=1,2,\cdots,n)$ 满足条件

$$\begin{gathered} \lambda_1 = \lambda_2 = \cdots = \lambda_m \\ |\lambda_1| > |\lambda_{m+1}| \geqslant |\lambda_{m+2}| \geqslant \cdots \geqslant |\lambda_n| \end{gathered} \tag{3.10}$$

但 A 仍有 n 个线性无关的特征向量。这时，由式(3.5)可看出，只要 $\alpha_1, \alpha_2, \cdots, \alpha_m$ 不全为零，则当 $\lambda_1 > 0$ 时，有

$$\lim_{k\to\infty} y_k = \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m}{\| \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m \|}$$

当 $\lambda_1 < 0$ 且 k 足够大时，有

$$y_k \approx (-1)^k \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m}{\| \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m \|}$$

由于向量 $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m$ 仍是 A 的属于 λ_1 的特征向量，故使用幂法迭代格式(3.7)和(3.9)，当 k 足够大时，y_k 都可以近似地作为 A 的属于 λ_1 的特征向量，而两种迭代格式中的 β_k 都收敛于 λ_1。

例 1 求矩阵

$$A = \begin{bmatrix} 6 & -12 & 6 \\ -21 & -3 & 24 \\ -12 & -12 & 51 \end{bmatrix}$$

按模最大的特征值 λ_1 和属于 λ_1 的特征向量，要求 $|\beta_k - \beta_{k-1}| / |\beta_k| \leqslant 0.000\,1$。

解 采用幂法迭代格式(3.9)进行计算。计算结果见表 3-1，表中的数字只写到小数点后第四位。

表 3-1 例 1 计算结果

k	u_k^T			y_k^T			β_k
0	1.000 0	0.000 0	0.000 0	1.000 0	0.000 0	0.000 0	
1	6.000 0	−21.000 0	−12.000 0	0.285 7	−1.000 0	−0.571 4	6.000 0
2	10.285 7	−16.714 3	−20.571 4	0.500 0	−0.812 5	−1.000 0	16.714 3
3	6.750 0	−32.062 5	−47.250 0	0.142 9	−0.678 6	−1.000 0	47.250 0
4	3.000 0	−24.964 3	−44.571 4	0.067 3	−0.560 1	−1.000 0	44.571 4
5	1.125 0	−23.733 2	−45.086 5	0.025 0	−0.526 4	−1.000 0	45.086 5
6	0.466 4	−22.944 8	−44.982 7	0.010 4	−0.510 1	−1.000 0	44.982 7
7	0.183 2	−22.687 5	−45.003 4	0.004 1	−0.504 1	−1.000 0	45.003 4
8	0.074 0	−22.573 1	−44.999 3	0.001 6	−0.501 6	−1.000 0	44.999 3
9	0.029 4	−22.529 6	−45.000 1				45.000 1

因 β_9 已满足条件，故 $\boldsymbol{A}$ 的模为最大的特征值是 $\lambda_1 \approx 45.0001$，相应的特征向量为

$$\boldsymbol{x}_1 \approx (0.0016,\ -0.5016,\ -1)^{\mathrm{T}}$$

[精确结果是 $\lambda_1 = 45$，$\boldsymbol{x}_1 = (0, -0.5, -1)^{\mathrm{T}}$]

对幂法这里只限于讨论特征值满足式(3.1)或(3.10)的两种情况。如果在幂法迭代格式(3.7)和(3.9)的迭代过程中，当 k 逐渐增大时，$\boldsymbol{y}_k$ 总是波动不定，说明矩阵 $\boldsymbol{A}$ 的特征值不满足式(3.1)和(3.10)。这时最好改用其他方法计算 $\boldsymbol{A}$ 的特征值和特征向量。

幂法的收敛速度与比值 $\left|\frac{\lambda_2}{\lambda_1}\right|$[式(3.1)的情况]或 $\left|\frac{\lambda_{m+1}}{\lambda_1}\right|$[式(3.10)的情况]有关，比值越小，收敛速度越快。此外，当矩阵 $\boldsymbol{A}$ 没有 n 个线性无关的特征向量，幂法仍然可以使用，但收敛速度特别慢，应改用其他方法。

3.1.2 反幂法

设 $n \times n$ 实矩阵 $\boldsymbol{A}$ 非奇异，且具有 n 个线性无关的特征向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，其特征值满足

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_{n-1}| > |\lambda_n|$$

其中 $\boldsymbol{A}\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i (i = 1, 2, \cdots, n)$，现在要计算 $\boldsymbol{A}$ 的按模最小的特征值 λ_n 以及相应的特征向量。

因 $\boldsymbol{A}$ 非奇异，故 $\lambda_i \neq 0 (i = 1, 2, \cdots, n)$。由

$$\boldsymbol{A}\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$$

得

$$\boldsymbol{A}^{-1}\boldsymbol{x}_i = \frac{1}{\lambda_i}\boldsymbol{x}_i \quad (i = 1, 2, \cdots, n)$$

所以，$\frac{1}{\lambda_n}$ 是矩阵 $\boldsymbol{A}^{-1}$ 的按模最大的特征值，$\boldsymbol{x}_n$ 是 $\boldsymbol{A}^{-1}$ 的属于 $\frac{1}{\lambda_n}$ 的特征向量。于是，对矩阵 $\boldsymbol{A}^{-1}$ 使用幂法迭代格式(3.7)或(3.9)就可求出 $\frac{1}{\lambda_n}$（从而求出 λ_n）以及相应的特征向量。这里使用幂法迭代格式(3.7)，得出如下的反幂法迭代格式：

$$
\begin{cases}
\text{任取非零向量 } \boldsymbol{u}_0 \in \mathbf{R}^n \\
\eta_{k-1} = \sqrt{\boldsymbol{u}_{k-1}^{\mathrm{T}}\boldsymbol{u}_{k-1}} \\
\boldsymbol{y}_{k-1} = \boldsymbol{u}_{k-1}/\eta_{k-1} \\
\boldsymbol{A}\boldsymbol{u}_k = \boldsymbol{y}_{k-1} \\
\beta_k = \boldsymbol{y}_{k-1}^{\mathrm{T}}\boldsymbol{u}_k \\
\qquad (k = 1,2,\cdots)
\end{cases}
\tag{3.11}
$$

每迭代一次都要求解一次线性方程组 $\boldsymbol{A}\boldsymbol{u}_k=\boldsymbol{y}_{k-1}$。当 k 足够大时，$\lambda_n\approx\frac{1}{\beta_k}$，$\boldsymbol{y}_{k-1}$ 可近似地作为矩阵 $\boldsymbol{A}$ 的属于 λ_n 的特征向量。比值 $\left|\frac{\lambda_n}{\lambda_{n-1}}\right|$ 越小，收敛得越快。

例 2 用反幂法求矩阵

$$
\boldsymbol{A}=\begin{bmatrix} 6 & -12 & 6 \\ -21 & -3 & 24 \\ -12 & -12 & 51 \end{bmatrix}
$$

的按模最小的特征值和相应的特征向量，要求 $|\beta_k^{-1}-\beta_{k-1}^{-1}|/|\beta_k^{-1}|\leqslant 0.005$。

解 使用反幂法迭代公式(3.11)，计算结果见表 3-2，表中的数字只写到小数点后第四位。

表 3-2 例 2 计算结果

k	$\boldsymbol{u}_k^{\mathrm{T}}$			$\boldsymbol{y}_k^{\mathrm{T}}$			β_k
0	1.000 0	1.000 0	1.000 0	0.577 4	0.577 4	0.577 4	
1	−0.032 1	−0.070 6	−0.012 8	−0.408 2	−0.898 1	−0.163 3	−0.066 7
2	0.068 0	0.084 4	0.008 5	0.627 8	0.778 4	0.078 1	−0.104 9
3	−0.066 4	−0.104 9	−0.038 8	−0.510 5	−0.806 5	−0.298 1	−0.126 4
4	0.058 2	0.085 6	0.028 0	0.542 4	0.798 5	0.261 0	−0.107 1
5	−0.059 5	−0.090 0	−0.030 1	−0.531 4	−0.803 5	−0.268 4	−0.112 0
6	0.059 4	0.088 8	0.029 6	0.535 9	0.800 9	0.267 1	−0.110 9
7	−0.059 4	−0.089 2	−0.029 7				−0.111 2

β_7 已满足精度要求。所以，$\boldsymbol{A}$ 的按模最小的特征值为 $\lambda_3\approx\frac{1}{\beta_7}=-8.992\,8$，相应的特征向量为 $\boldsymbol{x}_3\approx(0.535\,9,0.800\,9,0.267\,1)^{\mathrm{T}}$。

在实际计算中常采用带原点平移的反幂法求矩阵 $\boldsymbol{A}$ 的某个特征值 λ_s。此法的依据是：若 λ 是矩阵 $\boldsymbol{A}$ 的特征值，则 $\lambda-p$ 是矩阵 $\boldsymbol{A}-p\boldsymbol{I}$ 的特征值，其中 $\boldsymbol{I}$ 是单位矩阵；反之，若 $\lambda-p$ 是矩阵 $\boldsymbol{A}-p\boldsymbol{I}$ 的特征值，则 λ 是矩阵 $\boldsymbol{A}$ 的特征值，而特征向量是相同的。事实上，

$$
\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}
$$

与

$$
(\boldsymbol{A}-p\boldsymbol{I})\boldsymbol{x}=(\lambda-p)\boldsymbol{x}
$$

可互为因果，其中 $\boldsymbol{x}$ 是非零向量。

设已知数 μ 是 $n\times n$ 矩阵 $\boldsymbol{A}$ 的某个特征值 λ_s 的近似值，并满足

$$0<|\lambda_s-\mu|<|\lambda_i-\mu|,\quad 1\leqslant i\leqslant n, i\neq s$$

于是，对矩阵 $\boldsymbol{A}-\mu\boldsymbol{I}$ 实行反幂法迭代，就可求出 $\boldsymbol{A}$ 的特征值 λ_s 及其相应的特征向量。其中 μ 称为平移量。具体说，就是把反幂法迭代格式(3.11)中的 $\boldsymbol{A}\boldsymbol{u}_k=\boldsymbol{y}_{k-1}$ 改为

$$(\boldsymbol{A}-\mu\boldsymbol{I})\boldsymbol{u}_k=\boldsymbol{y}_{k-1}$$

当 k 足够大时，取 $\lambda_s\approx\beta_k^{-1}+\mu$，其相应的特征向量近似地等于 $\boldsymbol{y}_{k-1}$。

由于幂法和反幂法的迭代是否收敛依赖于特征值的分布情况，因此实际使用时很不方便，特别是不适合于自动计算。只在矩阵阶数非常高，无法利用其他更有效的算法时，才用幂法计算按模最大的特征值和相应的特征向量，而用反幂法计算按模最小的特征值和相应的特征向量。

3.2 Jacobi 方法

Jacobi 方法是用于求实对称矩阵的全部特征值和特征向量的一种方法。

对于一个实对称矩阵 $\boldsymbol{A}=[a_{ij}]_{n\times n}$，一定存在正交矩阵 $\boldsymbol{U}$，使

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}=\boldsymbol{D}$$

其中 $\boldsymbol{D}$ 是对角矩阵，其主对角线元素 $\lambda_j(j=1,2,\cdots,n)$ 是矩阵 $\boldsymbol{A}$ 的特征值，而正交矩阵 $\boldsymbol{U}$ 的第 j 列就是 $\boldsymbol{A}$ 的属于 λ_j 的特征向量。Jacobi 方法就是用平面旋转矩阵对矩阵 $\boldsymbol{A}$ 作正交相似变换把 $\boldsymbol{A}$ 化为对角矩阵，从而求出 $\boldsymbol{A}$ 的特征值和特征向量。

设有实对称矩阵 $\boldsymbol{A}=[a_{ij}]_{n\times n}$。它的一对非主对角线元素 $a_{pq}=a_{qp}$，且不为零($p\neq q$)。取 $n\times n$ 的正交矩阵 $\boldsymbol{U}_{pq}$ 为

$$U_{pq}=\begin{bmatrix}
1 & & & \vdots & & & & & & & \\
 & \ddots & & \vdots & & & & & & & \\
 & & 1 & \vdots & & & & & & & \\
\cdots & \cdots & \cdots & \cos\varphi & \cdots & \cdots & \cdots & -\sin\varphi & \cdots & \cdots & \cdots \\
 & & & \vdots & 1 & & & \vdots & & & \\
 & & & \vdots & & \ddots & & \vdots & & & \\
 & & & \vdots & & & 1 & \vdots & & & \\
\cdots & \cdots & \cdots & \sin\varphi & \cdots & \cdots & \cdots & \cos\varphi & \cdots & \cdots & \cdots \\
 & & & \vdots & & & & \vdots & 1 & & \\
 & & & \vdots & & & & \vdots & & \ddots & \\
 & & & \vdots & & & & \vdots & & & 1
\end{bmatrix}\begin{matrix} \\ \\ \\ \text{第 } p \text{ 行} \\ \\ \\ \\ \text{第 } q \text{ 行} \\ \\ \\ \\ \end{matrix}\qquad(3.12)$$

第 p 列　　　第 q 列

$\boldsymbol{U}_{pq}$ 的主对角线元素中，$u_{pp}=u_{qq}=\cos\varphi$，其余为 1；$\boldsymbol{U}_{pq}$ 的非主对角线元素中，$u_{pq}=-\sin\varphi$，$u_{qp}=\sin\varphi$，其余为零。$\boldsymbol{U}_{pq}$ 是 n 维空间中的二维坐标旋转变换矩阵。设 $\boldsymbol{x}\in\mathbf{R}^n$，则 $\boldsymbol{U}_{pq}\boldsymbol{x}$ 相当于将坐标轴 Ox_p 和 Ox_q 在 x_p, x_q 所在平面旋转了一个角度 φ，其他坐标轴保持不变，故称式(3.12)的 $\boldsymbol{U}_{pq}$ 为平面旋转矩阵。用 $\boldsymbol{U}_{pq}$ 对 $\boldsymbol{A}$ 作正交相似变换，得到矩阵 $\boldsymbol{A}_1$，即

$$\boldsymbol{A}_1=\boldsymbol{U}_{pq}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}_{pq}=[a_{ij}^{(1)}]_{n\times n}$$

显然，$\boldsymbol{A}_1$ 仍是实对称矩阵，$\boldsymbol{A}_1$ 与 $\boldsymbol{A}$ 的特征值完全相同。通过直接计算，知

$$\begin{cases} a_{pp}^{(1)} = a_{pp}\cos^2\varphi + a_{qq}\sin^2\varphi + 2a_{pq}\cos\varphi\sin\varphi \\ a_{qq}^{(1)} = a_{pp}\sin^2\varphi + a_{qq}\cos^2\varphi - 2a_{pq}\cos\varphi\sin\varphi \\ \left.\begin{array}{l} a_{pi}^{(1)} = a_{ip}^{(1)} = a_{pi}\cos\varphi + a_{qi}\sin\varphi \\ a_{qi}^{(1)} = a_{iq}^{(1)} = -a_{pi}\sin\varphi + a_{qi}\cos\varphi \end{array}\right\} \quad i \neq p,q \\ a_{ij}^{(1)} = a_{ji}^{(1)} = a_{ij}, \quad i,j \neq p,q \\ a_{pq}^{(1)} = a_{qp}^{(1)} = \dfrac{1}{2}(a_{qq} - a_{pp})\sin 2\varphi + a_{pq}\cos 2\varphi \end{cases} \tag{3.13}$$

可见,变换的结果,矩阵 $\boldsymbol{A}_1$ 的第 p 行、第 p 列和第 q 行、第 q 列的元素发生了变化,其余元素不变。特别地,当取 φ 满足关系式

$$\cot 2\varphi = \frac{a_{pp} - a_{qq}}{2a_{pq}} \tag{3.14}$$

时,可以得到 $a_{pq}^{(1)} = a_{qp}^{(1)} = 0$。也就是说,用平面旋转变换矩阵 $\boldsymbol{U}_{pq}$ 对 $\boldsymbol{A}$ 进行正交相似变换,可以将 $\boldsymbol{A}$ 的两个非主对角线元素 a_{pq} 和 a_{qp} 化为零。

求实对称矩阵 $\boldsymbol{A}$ 的特征值和特征向量是一个迭代过程,其迭代步骤如下:

(1) 在 $\boldsymbol{A}$ 的非主对角线元素中,找出按模最大的元素 a_{pq}。

(2) 由等式(3.14)计算 $\cot 2\varphi$,并由此求出 $\sin\varphi$,$\cos\varphi$ 以及相应的平面旋转矩阵 $\boldsymbol{U}_{pq}$。

(3) 按照公式(3.13)计算矩阵 $\boldsymbol{A}_1$ 的元素 $a_{ij}^{(1)}$。

(4) 若 $\max\limits_{i<j}|a_{ij}^{(1)}| < \varepsilon$(允许误差),则停止计算,所求特征值为 $\lambda_i \approx a_{ii}^{(1)}$ $(i=1,2,\cdots,n)$;否则,令 $\boldsymbol{A}=\boldsymbol{A}_1$,重复执行(1),(2),(3),(4)。

当条件 $\max\limits_{i<j}|a_{ij}^{(1)}| < \varepsilon$ 满足时,$\boldsymbol{A}_1$ 的所有非主对角线元素在所给精度要求下近似等于零,$\boldsymbol{A}_1$ 几乎是一个对角矩阵。因此,可取 $\boldsymbol{A}_1$ 的主对角线元素作为 $\boldsymbol{A}$ 的特征值的近似值,即

$$\lambda_i \approx a_{ii}^{(1)} \quad (i = 1,2,\cdots,n)$$

设经过 N 次迭代,上述条件得到满足,又记第 k 次迭代所得的平面旋转矩阵为 $\boldsymbol{U}_{p_k q_k}$,那么,经过 N 次迭代所得的矩阵为

$$\boldsymbol{A}_1 = \boldsymbol{U}_{p_N q_N}^{\mathrm{T}} \cdots \boldsymbol{U}_{p_2 q_2}^{\mathrm{T}} \boldsymbol{U}_{p_1 q_1}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{U}_{p_1 q_1} \boldsymbol{U}_{p_2 q_2} \cdots \boldsymbol{U}_{p_N q_N}$$

记

$$\boldsymbol{U} = \boldsymbol{U}_{p_1 q_1} \boldsymbol{U}_{p_2 q_2} \cdots \boldsymbol{U}_{p_N q_N}$$

则 $\boldsymbol{U}$ 是正交矩阵,并且有

$$\boldsymbol{A}_1 = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{U}$$

因为 $\boldsymbol{A}_1$ 被看做是对角矩阵,所以矩阵 $\boldsymbol{U}$ 的第 j 列就是 $\boldsymbol{A}$ 的属于特征值 $\lambda_j \approx a_{jj}^{(1)}$ 的近似特征向量,并且所有的特征向量都是正交规范化的。

在旋转变换中可以逐步形成 $\boldsymbol{U}$,而不必保存每一次的变换矩阵 $\boldsymbol{U}_{p_k q_k}$。记

$$\boldsymbol{R}_0 = \boldsymbol{I}$$

令

$$\boldsymbol{R}_k = \boldsymbol{R}_{k-1} \boldsymbol{U}_{p_k q_k} \quad (k = 1,2,\cdots,N) \tag{3.15}$$

则

$$\boldsymbol{U} = \boldsymbol{R}_N$$

根据式(3.15),计算矩阵 $\boldsymbol{R}_k$ 的元素 $r_{ij}^{(k)}$ 的公式为

$$
\begin{cases}
r_{ip_k}^{(k)} = r_{ip_k}^{(k-1)}\cos\varphi + r_{iq_k}^{(k-1)}\sin\varphi \\
r_{iq_k}^{(k)} = -r_{ip_k}^{(k-1)}\sin\varphi + r_{iq_k}^{(k-1)}\cos\varphi \\
r_{ij}^{(k)} = r_{ij}^{(k-1)},\quad j \neq p_k, q_k \\
\qquad (i = 1,2,3,\cdots,n)
\end{cases}
\tag{3.16}
$$

用 Jacobi 方法计算 $n\times n$ 矩阵 $\boldsymbol{A}$ 的特征值和特征向量，每迭代一次都要按公式(3.13)、(3.14)和(3.16)计算一次。计算量主要是乘法的次数，约为 $8n$ 次。

关于 Jacobi 方法的收敛性，有以下的定理。

定理 3.1　设 $\boldsymbol{A}=[a_{ij}]_{n\times n}$ 是实对称矩阵，由 Jacobi 方法的第 k 次迭代得到的矩阵记为 $\boldsymbol{A}_k=[a_{ij}^{(k)}]_{n\times n}$，又记

$$
\eta_k = \sum_{\substack{i,j=1 \\ i\neq j}}^{n} (a_{ij}^{(k)})^2
$$

则有 $\lim\limits_{k\to\infty}\eta_k=0$ 成立。

证　令

$$
\delta_k = \sum_{i=1}^{n} (a_{ii}^{(k)})^2
$$

由式(3.13)经过计算，并注意到

$$
\frac{1}{2}(a_{qq}^{(k)} - a_{pp}^{(k)})\sin 2\varphi + a_{pq}^{(k)}\cos 2\varphi = 0
$$

可得

$$
(a_{pp}^{(k+1)})^2 + (a_{qq}^{(k+1)})^2 = (a_{pp}^{(k)})^2 + (a_{qq}^{(k)})^2 + 2(a_{pq}^{(k)})^2
$$

又由于

$$
a_{ii}^{(k+1)} = a_{ii}^{(k)},\quad i \neq p,q
$$

所以

$$
\delta_{k+1} = \delta_k + 2(a_{pq}^{(k)})^2
$$

因实对称矩阵经过正交相似变换后，其元素的平方和不变，即

$$
\eta_{k+1} + \delta_{k+1} = \eta_k + \delta_k
$$

故有

$$
\eta_{k+1} = \eta_k - (\delta_{k+1} - \delta_k) = \eta_k - 2(a_{pq}^{(k)})^2
$$

又因

$$
|a_{pq}^{(k)}| = \max_{i<j} |a_{ij}^{(k)}|
$$

故得

$$
\eta_k \leqslant n(n-1)(a_{pq}^{(k)})^2
$$

从而有

$$
\eta_{k+1} = \eta_k - 2(a_{pq}^{(k)})^2 \leqslant \eta_k - \frac{2}{n(n-1)}\eta_k = \left[1 - \frac{2}{n(n-1)}\right]\eta_k
\tag{3.17}
$$

反复使用式(3.17)，得

$$
\eta_{k+1} \leqslant \left[1 - \frac{2}{n(n-1)}\right]^{k+1}\eta_0
$$

其中 η_0 是矩阵 $\boldsymbol{A}$ 的非主对角线元素平方之和。

由于

$$0 \leqslant 1 - \frac{2}{n(n-1)} < 1$$

所以有$\lim\limits_{k \to \infty} \eta_k = 0$。

证毕。

前面叙述的 Jacobi 方法又称为经典的 Jacobi 方法。它每次迭代都是把按模最大的非主对角线元素 a_{pq} 作为消元对象。不论实对称矩阵 $\boldsymbol{A}$ 的特征值如何分布，经典的 Jacobi 方法总是收敛的，而且当 $\boldsymbol{A}$ 的阶数不太高时，收敛速度还比较快。此外，这个方法具有较强的数值稳定性，求得的结果精度一般都比较高，特别是求得的特征向量正交性很好，这是其他方法所不如的。经典的 Jacobi 方法的缺点是，不能有效地利用矩阵的各种特殊形状（例如带状或稀疏等）以节省工作量。这是因为它的迭代过程中一般都会破坏原矩阵的特殊形状。还有其他一些缺点。例如，绝对值较小的特征值精度差一些；由于每迭代一次都要在非主对角线元素中寻找模数最大的元素，因而比较费时间。

为节省计算时间，提高精度，实际使用的 Jacobi 方法常采取以下的措施。

（1）按行循环消元，即在迭代过程中，旋转矩阵 $\boldsymbol{U}_{pq}$ 中的(p,q)按照下列次序选取：

$$(1,2),(1,3),\cdots,(1,n),(2,3),(2,4),\cdots,(2,n),\cdots,(n-1,n)$$

如果$|a_{pq}|$与主对角线元素的平方和之比很小，这一步就可以跳过。从$(1,2)$到$(n-1,n)$称为一轮。也可以采用变容限循环消元法，就是在每一轮消元时，都给定一个控制量 ε，称为消元容限，如果$|a_{pq}|<\varepsilon$ 就跳过这一步。容限 ε 的值逐渐减小，最后可取接近计算机所能表示的最小正数作为容限。

（2）计算旋转矩阵中的 $\sin\varphi$ 和 $\cos\varphi$ 对于计算结果的精确度有很大影响。为了使得计算 $\sin\varphi$ 和 $\cos\varphi$ 尽可能准确，可采用下面的公式：

$$t = \frac{2a_{pq}}{a_{pp} - a_{qq}}, \quad z = \frac{a_{pp} - a_{qq}}{2a_{pq}}$$

$$\cos 2\varphi = \begin{cases} (1+t^2)^{-\frac{1}{2}}, & |t| < 1 \\ |z|(1+z^2)^{-\frac{1}{2}}, & |t| \geqslant 1 \end{cases}$$

$$\sin 2\varphi = \begin{cases} t(1+t^2)^{-\frac{1}{2}}, & |t| < 1 \\ \operatorname{sgn} z \cdot (1+z^2)^{-\frac{1}{2}}, & |t| \geqslant 1 \end{cases}$$

$$\cos\varphi = \left[\frac{1}{2}(1+\cos 2\varphi)\right]^{\frac{1}{2}}$$

$$\sin\varphi = \frac{\sin 2\varphi}{2\cos\varphi}$$

3.3 QR 方法

3.3.1 矩阵的 QR 分解

把矩阵 $\boldsymbol{A}$ 分解为一个正交矩阵 $\boldsymbol{Q}$ 与一个上三角矩阵 $\boldsymbol{R}$ 的乘积，称为矩阵 $\boldsymbol{A}$ 的正交三角分解，简称 QR 分解。下面讨论对矩阵作 QR 分解的可能性和分解方法。

设 $v\in\mathbf{R}^n$ 是单位向量，即 $v^Tv=1$，令

$$H = I - 2vv^T \tag{3.18}$$

其中 I 是 $n\times n$ 单位矩阵。易知，由式(3.18)确定的矩阵 H 满足

$$H^T = H$$

$$HH^T = (I-2vv^T)(I-2vv^T) = I - 4vv^T + 4v(v^Tv)v^T = I$$

因此，H 是对称正交矩阵。

由式(3.18)确定的矩阵 H 称为 Householder（豪斯荷尔德）矩阵，又称为镜面映射矩阵。由式(3.18)看出，Householder 矩阵 H 由单位向量 v 唯一确定。

引理 3.1　设有非零向量 $s\in\mathbf{R}^n$ 和单位向量 $e\in\mathbf{R}^n$，必存在 Householder 矩阵 H，使得

$$Hs = \alpha e$$

其中 α 是实数，并且 $|\alpha|=\sqrt{s^Ts}$。

证　取单位向量

$$v = \frac{1}{\rho}(s-\alpha e) \tag{3.19}$$

其中 $\rho=\sqrt{(s-\alpha e)^T(s-\alpha e)}$。把式(3.19)代入式(3.18)形成 Householder 矩阵 H，则有

$$Hs = (I-2vv^T)s = s - \frac{2}{\rho}v(s^T-\alpha e^T)s = s - \frac{2}{\rho}v(\alpha^2 - \alpha e^Ts)$$

因

$$\rho^2 = (s-\alpha e)^T(s-\alpha e) = 2(\alpha^2 - \alpha e^Ts)$$

故

$$Hs = s - \rho v = \alpha e$$

证毕。

定理 3.2　任何 $n\times n$ 实矩阵 A 总可以分解为一个正交矩阵 Q 与一个上三角矩阵 R 的乘积。

证　用构造法证明。

设 $A=[a_{ij}]_{n\times n}$，$e_r=(0,\cdots,0,1,0,\cdots,0)^T$ 是 n 维基本单位向量，它的第 r 个分量为 1，其余分量为零。

设 $a_{i1}(i=2,3,\cdots,n)$ 不全为零，令

$$s_1=(a_{11},\cdots,a_{n1})^T$$

$$c_1=-\mathrm{sgn}\,(a_{11})\sqrt{s_1^Ts_1}\quad（若\ a_{11}=0，则取\ c_1=\sqrt{s_1^Ts_1}）$$

$$u_1=s_1-c_1e_1$$

构成 Householder 矩阵

$$H_1 = I - 2u_1u_1^T/(u_1^Tu_1)$$

根据引理 3.1，有

$$H_1s_1 = c_1e_1 = (c_1,0,\cdots,0)^T$$

因而有

$$A_2 = H_1A = \begin{bmatrix} c_1 & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}$$

若 $a_{i1}(i=2,3,\cdots,n)$全为零，则取 $\boldsymbol{H}_1=\boldsymbol{I}$，并且

$$\boldsymbol{A}_2=\boldsymbol{H}_1\boldsymbol{A}=\boldsymbol{A}$$

设 $a_{i2}^{(2)}(i=3,4,\cdots,n)$不全为零，令

$$\boldsymbol{s}_2=(0,a_{22}^{(2)},\cdots,a_{n2}^{(2)})^{\mathrm{T}}$$

$$c_2=-\operatorname{sgn}(a_{22}^{(2)})\sqrt{\boldsymbol{s}_2^{\mathrm{T}}\boldsymbol{s}_2}\quad(\text{若 } a_{22}^{(2)}=0\text{，则取 } c_2=\sqrt{\boldsymbol{s}_2^{\mathrm{T}}\boldsymbol{s}_2})$$

$$\boldsymbol{u}_2=\boldsymbol{s}_2-c_2\boldsymbol{e}_2$$

构成 Householder 矩阵

$$\boldsymbol{H}_2=\boldsymbol{I}-2\boldsymbol{u}_2\boldsymbol{u}_2^{\mathrm{T}}/(\boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{u}_2)=\begin{bmatrix}1&\boldsymbol{0}\\\boldsymbol{0}&\boldsymbol{W}_1\end{bmatrix}$$

其中 $\boldsymbol{W}_1$ 是$(n-1)\times(n-1)$矩阵，可得

$$\boldsymbol{H}_2\boldsymbol{s}_2=c_2\boldsymbol{e}_2=(0,c_2,0,\cdots,0)^{\mathrm{T}}$$

$$\boldsymbol{A}_3=\boldsymbol{H}_2\boldsymbol{A}_2=\begin{bmatrix}c_1&a_{12}^{(2)}&a_{13}^{(2)}&\cdots&a_{1n}^{(2)}\\0&c_2&a_{23}^{(3)}&\cdots&a_{2n}^{(3)}\\\vdots&0&\vdots&&\vdots\\\vdots&\vdots&\vdots&&\vdots\\0&0&a_{n3}^{(3)}&\cdots&a_{nn}^{(3)}\end{bmatrix}$$

若 $a_{i2}^{(2)}(i=3,4,\cdots,n)$全为零，则取 $\boldsymbol{H}_2=\boldsymbol{I}$，并且

$$\boldsymbol{A}_3=\boldsymbol{H}_2\boldsymbol{A}_2=\boldsymbol{A}_2$$

一般地，设按照上述方法已得到矩阵 $\boldsymbol{A}_r(r\geqslant 2)$，又设 $a_{ir}^{(r)}(i=r+1,r+2,\cdots,n)$不全为零，则令

$$\boldsymbol{s}_r=(0,\cdots,0,a_{rr}^{(r)},\cdots,a_{nr}^{(r)})^{\mathrm{T}}$$

$$c_r=-\operatorname{sgn}(a_{rr}^{(r)})\sqrt{\boldsymbol{s}_r^{\mathrm{T}}\boldsymbol{s}_r}\quad(\text{若 } a_{rr}^{(r)}=0\text{，则取 } c_r=\sqrt{\boldsymbol{s}_r^{\mathrm{T}}\boldsymbol{s}_r})$$

$$\boldsymbol{u}_r=\boldsymbol{s}_r-c_r\boldsymbol{e}_r$$

构成 Householder 矩阵

$$\boldsymbol{H}_r=\boldsymbol{I}-2\boldsymbol{u}_r\boldsymbol{u}_r^{\mathrm{T}}/(\boldsymbol{u}_r^{\mathrm{T}}\boldsymbol{u}_r)=\begin{bmatrix}\boldsymbol{I}_{r-1}&\boldsymbol{0}\\\boldsymbol{0}&\boldsymbol{W}_{r-1}\end{bmatrix}$$

其中 $\boldsymbol{I}_{r-1}$是$(r-1)\times(r-1)$单位矩阵，$\boldsymbol{W}_{r-1}$是$(n-r+1)\times(n-r+1)$矩阵，可得

$$\boldsymbol{H}_r\boldsymbol{s}_r=c_r\boldsymbol{e}_r=(0,\cdots,0,c_r,0,\cdots,0)^{\mathrm{T}}$$

$$\boldsymbol{A}_{r+1}=\boldsymbol{H}_r\boldsymbol{A}_r=\begin{bmatrix}c_1&\cdots&a_{1r}^{(2)}&a_{1,r+1}^{(2)}&\cdots&a_{1n}^{(2)}\\&\ddots&\vdots&\vdots&&\vdots\\&&c_r&a_{r,r+1}^{(r+1)}&\cdots&a_{rn}^{(r+1)}\\&&&a_{r+1,r+1}^{(r+1)}&\cdots&a_{r+1,n}^{(r+1)}\\&&&\vdots&&\vdots\\&&&a_{n,r+1}^{(r+1)}&\cdots&a_{nn}^{(r+1)}\end{bmatrix}$$

如果 $a_{ir}^{(r)}(i=r+1,r+2,\cdots,n)$全为零，则取 $\boldsymbol{H}_r=\boldsymbol{I}$，这时有 $\boldsymbol{A}_{r+1}=\boldsymbol{H}_r\boldsymbol{A}_r=\boldsymbol{A}_r$。

于是，当 $r=n-1$ 时，就得到 Householder 矩阵（或者是单位矩阵）$\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_{n-1}$，使得

$$\boldsymbol{A}_n=\boldsymbol{H}_{n-1}\boldsymbol{H}_{n-2}\cdots\boldsymbol{H}_1\boldsymbol{A}\tag{3.20}$$

是一个上三角矩阵。由式(3.20)得

$$A = (H_{n-1}H_{n-2}\cdots H_1)^{-1}A_n = H_1H_2\cdots H_{n-1}A_n$$

记

$$Q = H_1H_2\cdots H_{n-1}, \quad R = A_n$$

则有

$$A = QR$$

其中 Q 是正交矩阵，R 是上三角矩阵。

证毕。

对 $n\times n$ 实矩阵 A 作 QR 分解，实际计算时不必具体形成矩阵 H_i，并且可避免矩阵与矩阵相乘。具体算法如下：

记 $A_1=A$，并记 $A_r=[a_{ij}^{(r)}]_{n\times n}$，令 $Q_1=I$（n 阶单位矩阵）。

对于 $r=1,2,\cdots,n-1$ 执行

（1）若 $a_{ir}^{(r)}(i=r+1,r+2,\cdots,n)$全为零，则令

$$Q_{r+1} = Q_r, \quad A_{r+1} = A_r$$

转(5)；否则转(2)。

（2）计算

$$d_r = \sqrt{\sum_{i=r}^{n}(a_{ir}^{(r)})^2}$$

$$c_r = -\operatorname{sgn}(a_{rr}^{(r)})\, d_r \quad （若\ a_{rr}^{(r)} = 0，则取\ c_r = d_r）$$

$$h_r = c_r^2 - c_r a_{rr}^{(r)}$$

（3）令 $u_r=(0,\cdots,0,a_{rr}^{(r)}-c_r,a_{r+1,r}^{(r)},\cdots,a_{nr}^{(r)})^{\mathrm{T}}\in \mathbf{R}^n$。

（4）计算

$$\omega_r = Q_r u_r$$

$$Q_{r+1} = Q_r - \omega_r u_r^{\mathrm{T}}/h_r$$

$$p_r = A_r^{\mathrm{T}} u_r/h_r$$

$$A_{r+1} = A_r - u_r p_r^{\mathrm{T}}$$

（5）继续。

当此算法执行完后就得到正交矩阵 $Q=Q_n$ 和上三角矩阵

$$R = A_n = \begin{bmatrix} c_1 & a_{12}^{(n)} & \cdots & a_{1,n-1}^{(n)} & a_{1n}^{(n)} \\ & \ddots & & \vdots & \vdots \\ & & \ddots & \vdots & \vdots \\ & & & c_{n-1} & a_{n-1,n}^{(n)} \\ & & & & a_{nn}^{(n)} \end{bmatrix}$$

且有 $A=QR$。

3.3.2　矩阵的拟上三角化

对实矩阵 $A=[a_{ij}]_{n\times n}$作相似变换化为拟上三角矩阵（又称上 Hessenberg 矩阵）$A^{(n-1)} = [\tilde{a}_{ij}]_{n\times n}$，其中 $i>j+1$ 时 $\tilde{a}_{ij}=0$，称为矩阵 A 的拟上三角化，其变换矩阵可采用 Householder 矩阵，变换的过程如下：

设 $a_{i1}(i=3,4,\cdots,n)$不全为零，令

$$s_1 = (0, a_{21}, \cdots, a_{n1})^{\mathrm{T}}$$
$$c_1 = -\operatorname{sgn}(a_{21}) \| s_1 \|_2 \quad (\text{若 } a_{21} = 0, \text{则取 } c_1 = \| s_1 \|_2)$$
$$u_1 = s_1 - c_1 e_2$$

构成 Householder 矩阵

$$H_1 = I - 2u_1u_1^{\mathrm{T}}/(u_1^{\mathrm{T}}u_1) = \begin{bmatrix} 1 & 0 \\ 0 & W_1 \end{bmatrix}$$

其中 W_1 是$(n-1)\times(n-1)$矩阵。根据引理 3.1，有

$$H_1 s_1 = c_1 e_2 = (0, c_1, 0, \cdots, 0)^{\mathrm{T}}$$

因而得

$$A^{(2)} = H_1 A H_1 = \begin{bmatrix} a_{11} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ c_1 & \vdots & & \vdots \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}$$

若 $a_{i1}(i=3,4,\cdots,n)$全为零，则取 $H_1=I$，并且

$$A^{(2)} = H_1 A H_1 = A$$

设 $a_{i2}^{(2)}(i=4,5,\cdots,n)$不全为零，令

$$s_2 = (0, 0, a_{32}^{(2)}, \cdots, a_{n2}^{(2)})^{\mathrm{T}}$$
$$c_2 = -\operatorname{sgn}(a_{32}^{(2)}) \| s_2 \|_2 \quad (\text{若 } a_{32}^{(2)} = 0, \text{则取 } c_2 = \| s_2 \|_2)$$
$$u_2 = s_2 - c_2 e_3$$

构成 Householder 矩阵

$$H_2 = I - 2u_2u_2^{\mathrm{T}}/(u_2^{\mathrm{T}}u_2) = \begin{bmatrix} I_2 & 0 \\ 0 & W_2 \end{bmatrix}$$

其中 I_2 是 2×2 单位矩阵，W_2 是$(n-2)\times(n-2)$矩阵。可得

$$H_2 s_2 = c_2 e_3 = (0, 0, c_2, 0, \cdots, 0)^{\mathrm{T}}$$

$$A^{(3)} = H_2 A^{(2)} H_2 = \begin{bmatrix} a_{11} & a_{12}^{(2)} & a_{13}^{(3)} & \cdots & a_{1n}^{(3)} \\ c_1 & a_{22}^{(2)} & \vdots & & \vdots \\ & c_2 & \vdots & & \vdots \\ & & \vdots & & \vdots \\ & & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} \end{bmatrix}$$

若 $a_{i2}^{(2)}(i=4,5,\cdots,n)$全为零，则取 $H_2=I$，并且

$$A^{(3)} = H_2 A^{(2)} H_2 = A^{(2)}$$

一般地，设按照上述方法已得到矩阵 $A^{(r)}(r\geqslant 2)$，又设 $a_{ir}^{(r)}(i=r+2,r+3,\cdots,n)$不全为零，则令

$$s_r = (0, \cdots, 0, a_{r+1,r}^{(r)}, \cdots, a_{nr}^{(r)})^{\mathrm{T}}$$
$$c_r = -\operatorname{sgn}(a_{r+1,r}^{(r)}) \| s_r \|_2 (\text{若 } a_{r+1,r}^{(r)} = 0, \text{则取 } c_r = \| s_r \|_2)$$
$$u_r = s_r - c_r e_{r+1}$$

构成 Householder 矩阵

$$\boldsymbol{H}_r = \boldsymbol{I} - 2\boldsymbol{u}_r\boldsymbol{u}_r^{\mathrm{T}}/(\boldsymbol{u}_r^{\mathrm{T}}\boldsymbol{u}_r) = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}_r \end{bmatrix}$$

其中 $\boldsymbol{I}_r$ 是 $r\times r$ 单位矩阵，$\boldsymbol{W}_r$ 是 $(n-r)\times(n-r)$ 矩阵，可得

$$\boldsymbol{H}_r\boldsymbol{s}_r = c_r\boldsymbol{e}_{r+1} = (0,\cdots,0,c_r,0,\cdots,0)^{\mathrm{T}}$$

$$\boldsymbol{A}^{(r+1)} = \boldsymbol{H}_r\boldsymbol{A}^{(r)}\boldsymbol{H}_r = \begin{bmatrix} a_{11} & \cdots & a_{1r}^{(r)} & a_{1,r+1}^{(r+1)} & \cdots & a_{1n}^{(r+1)} \\ c_1 & \ddots & \vdots & \vdots & & \vdots \\ & \ddots & a_{rr}^{(r)} & \vdots & & \vdots \\ & & c_r & \vdots & & \vdots \\ & & & a_{n,r+1}^{(r+1)} & \cdots & a_{nn}^{(r+1)} \end{bmatrix}$$

若 $a_{ir}^{(r)}(i=r+2,r+3,\cdots,n)$ 全为零，则取 $\boldsymbol{H}_r=\boldsymbol{I}$，并且有

$$\boldsymbol{A}^{(r+1)} = \boldsymbol{H}_r\boldsymbol{A}^{(r)}\boldsymbol{H}_r = \boldsymbol{A}^{(r)}$$

当 $r=n-2$ 时，就得拟上三角矩阵

$$\boldsymbol{A}^{(n-1)} = \boldsymbol{H}_{n-2}\cdots\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_{n-2} = \begin{bmatrix} a_{11} & \cdots & a_{1,n-2}^{(n-2)} & a_{1,n-1}^{(n-1)} & a_{1n}^{(n-1)} \\ c_1 & \ddots & \vdots & \vdots & \vdots \\ & \ddots & a_{n-2,n-2}^{(n-2)} & \vdots & \vdots \\ & & c_{n-2} & \vdots & \vdots \\ & & & a_{n,n-1}^{(n-1)} & a_{nn}^{(n-1)} \end{bmatrix} \tag{3.21}$$

因 $\boldsymbol{H}_r(r=1,2,\cdots,n-2)$ 是对称正交矩阵，故

$$\boldsymbol{P} = \boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_{n-2}$$

是正交矩阵，并且

$$\boldsymbol{P}^{\mathrm{T}} = (\boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_{n-2})^{\mathrm{T}} = \boldsymbol{H}_{n-2}\cdots\boldsymbol{H}_2\boldsymbol{H}_1$$

所以有

$$\boldsymbol{A}^{(n-1)} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}$$

因而 $\boldsymbol{A}^{(n-1)}$ 与 $\boldsymbol{A}$ 相似。

特别是当 $\boldsymbol{A}$ 为实对称矩阵时，$\boldsymbol{A}^{(n-1)}$ 也是实对称矩阵，因而 $\boldsymbol{A}^{(n-1)}$ 是对称三对角矩阵：

$$\boldsymbol{A}^{(n-1)} = \begin{bmatrix} a_{11} & c_1 & & & \\ c_1 & a_{22}^{(2)} & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-2} & a_{n-1,n-1}^{(n-1)} & a_{n,n-1}^{(n-1)} \\ & & & a_{n,n-1}^{(n-1)} & a_{nn}^{(n-1)} \end{bmatrix}$$

实际计算时不必具体形成矩阵 $\boldsymbol{H}_i$，并可避免矩阵与矩阵相乘。对实矩阵 $\boldsymbol{A}$ 的拟上三角化具体算法如下：

记 $\boldsymbol{A}^{(1)}=\boldsymbol{A}$，并记 $\boldsymbol{A}^{(r)}$ 的第 r 列至第 n 列的元素为 $a_{ij}^{(r)}(i=1,2,\cdots,n;j=r,r+1,\cdots,n)$。

对于 $r=1,2,\cdots,n-2$ 执行

(1) 若 $a_{ir}^{(r)}(i=r+2,r+3,\cdots,n)$ 全为零，则令 $\boldsymbol{A}^{(r+1)}=\boldsymbol{A}^{(r)}$，转(5)；否则转(2)。

(2) 计算

$$d_r = \sqrt{\sum_{i=r+1}^{n}(a_{ir}^{(r)})^2}$$

$$c_r = -\operatorname{sgn}(a_{r+1,r}^{(r)})\, d_r \quad (\text{若 } a_{r+1,r}^{(r)} = 0,\text{则取 } c_r = d_r)$$

$$h_r = c_r^2 - c_r a_{r+1,r}^{(r)}$$

(3) 令 $\boldsymbol{u}_r = (0,\cdots,0,a_{r+1,r}^{(r)} - c_r, a_{r+2,r}^{(r)},\cdots,a_{nr}^{(r)})^{\mathrm{T}} \in \mathbf{R}^n$。

(4) 计算

$$\begin{aligned}
\boldsymbol{p}_r &= \boldsymbol{A}^{(r)\mathrm{T}}\boldsymbol{u}_r/h_r \\
\boldsymbol{q}_r &= \boldsymbol{A}^{(r)}\boldsymbol{u}_r/h_r \\
t_r &= \boldsymbol{p}_r^{\mathrm{T}}\boldsymbol{u}_r/h_r \\
\boldsymbol{\omega}_r &= \boldsymbol{q}_r - t_r\boldsymbol{u}_r \\
\boldsymbol{A}^{(r+1)} &= \boldsymbol{A}^{(r)} - \boldsymbol{\omega}_r\boldsymbol{u}_r^{\mathrm{T}} - \boldsymbol{u}_r\boldsymbol{p}_r^{\mathrm{T}}
\end{aligned}$$

(5) 继续。

当此算法执行完后，就得到与原矩阵 $\boldsymbol{A}$ 相似的拟上三角矩阵 $\boldsymbol{A}^{(n-1)}$[参见式(3.21)]。

3.3.3 带双步位移的 QR 方法

QR 方法适用于计算一般实矩阵的全部特征值，尤其适用于计算中小型实矩阵的全部特征值。基本 QR 方法的迭代公式是

$$\begin{cases}
\boldsymbol{A}_1 = \boldsymbol{A} \in \mathbf{R}^{n\times n} \\
\boldsymbol{A}_k = \boldsymbol{Q}_k\boldsymbol{R}_k \quad (\text{对 } \boldsymbol{A}_k \text{ 作 QR 分解}) \\
\boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k \\
\qquad (k = 1,2,\cdots)
\end{cases} \tag{3.22}$$

由于

$$\boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k = \boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{Q}_k$$

所以，由迭代公式(3.22)产生的矩阵序列$\{\boldsymbol{A}_k\}$中的每个矩阵都与原矩阵 $\boldsymbol{A}$ 相似，因而任一矩阵 $\boldsymbol{A}_k$ 都与原矩阵 $\boldsymbol{A}$ 有相同的特征值。

关于 QR 方法的收敛性有这样的结论：如果矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的等模特征值中只有实特征值或复共轭特征值，则由 QR 方法(3.22)产生的矩阵序列$\{\boldsymbol{A}_k\}$本质上收敛于分块上三角矩阵(对角块以上的元素可能不收敛)，其对角块均为一阶和二阶子块，并且对角块中每一个一阶子块给出 $\boldsymbol{A}$ 的实特征值，每一个二阶子块给出 $\boldsymbol{A}$ 的一对复共轭特征值。特别是，当 $\boldsymbol{A}$ 为实对称矩阵时，QR 方法(3.22)产生的矩阵序列$\{\boldsymbol{A}_k\}$收敛于对角矩阵 $\boldsymbol{D}=\operatorname{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$，$\lambda_i(i=1,2,\cdots,n)$就是矩阵 $\boldsymbol{A}$ 的全部特征值。

为了减少计算量，一般先利用 Householder 矩阵对矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 作相似变换，把 $\boldsymbol{A}$ 化为拟上三角矩阵 $\boldsymbol{A}^{(n-1)}$，然后用 QR 方法计算 $\boldsymbol{A}^{(n-1)}$ 的全部特征值，而 $\boldsymbol{A}^{(n-1)}$ 的特征值就是 $\boldsymbol{A}$ 的特征值。

为了加速收敛，对基本 QR 方法(3.22)进行改进，改进为下面的带双步位移的 QR 方法：

$$\begin{cases}
\boldsymbol{A}_1 = \boldsymbol{A}^{(n-1)} \\
\boldsymbol{A}_k - s_1^{(k)}\boldsymbol{I} = \boldsymbol{Q}_k\boldsymbol{R}_k \quad (\text{对 } \boldsymbol{A}_k - s_1^{(k)}\boldsymbol{I} \text{ 作 QR 分解}) \\
\boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k + s_1^{(k)}\boldsymbol{I} \\
\boldsymbol{A}_{k+1} - s_2^{(k)}\boldsymbol{I} = \boldsymbol{Q}_{k+1}\boldsymbol{R}_{k+1} \quad (\text{对 } \boldsymbol{A}_{k+1} - s_2^{(k)}\boldsymbol{I} \text{ 作 QR 分解}) \\
\boldsymbol{A}_{k+2} = \boldsymbol{R}_{k+1}\boldsymbol{Q}_{k+1} + s_2^{(k)}\boldsymbol{I} \\
\qquad (k = 1,3,5,\cdots)
\end{cases} \tag{3.23}$$

其中 $s_1^{(k)}, s_2^{(k)}$ 是一对共轭复数或一对实数，称为位移量。

由迭代公式(3.23)可知

$$\boldsymbol{A}_{k+1} = \boldsymbol{Q}_k^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{Q}_k \tag{3.24}$$

$$\boldsymbol{A}_{k+2} = \boldsymbol{Q}_{k+1}^{\mathrm{T}} \boldsymbol{A}_{k+1} \boldsymbol{Q}_{k+1} = \boldsymbol{Q}_{k+1}^{\mathrm{T}} \boldsymbol{Q}_k^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{Q}_k \boldsymbol{Q}_{k+1} = \widetilde{\boldsymbol{Q}}_k^{\mathrm{T}} \boldsymbol{A}_k \widetilde{\boldsymbol{Q}}_k \tag{3.25}$$

其中 $\widetilde{\boldsymbol{Q}}_k = \boldsymbol{Q}_k \boldsymbol{Q}_{k+1}$ 是正交矩阵。由式(3.24)和式(3.25)可知，迭代公式(3.23)产生的矩阵序列中每个矩阵 $\boldsymbol{A}_k$ 均与矩阵 $\boldsymbol{A}_1$ 相似，都有与 $\boldsymbol{A}_1$ 相同的特征值。此外，只要 $\boldsymbol{A}_1$ 是拟上三角矩阵，则矩阵序列中的每个矩阵 $\boldsymbol{A}_k$ 都是拟上三角矩阵。引入位移量 $s_1^{(k)}$ 和 $s_2^{(k)}$ 可加速收敛，并选取 $\boldsymbol{A}_k$ 右下角二阶子阵

$$\boldsymbol{D}_k = \begin{bmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{bmatrix}$$

的两个特征值作为 $s_1^{(k)}$ 和 $s_2^{(k)}$。

为了避免复数运算，引入矩阵

$$\boldsymbol{M}_k = (\boldsymbol{A}_k - s_2^{(k)} \boldsymbol{I})(\boldsymbol{A}_k - s_1^{(k)} \boldsymbol{I}) = \boldsymbol{A}_k^2 - s\boldsymbol{A}_k + t\boldsymbol{I} \tag{3.26}$$

其中

$$s = s_1^{(k)} + s_2^{(k)} = a_{n-1,n-1}^{(k)} + a_{nn}^{(k)} \tag{3.27}$$

$$t = s_1^{(k)} s_2^{(k)} = \det \boldsymbol{D}_k \tag{3.28}$$

由于 s 和 t 都是实数，所以 $\boldsymbol{M}_k$ 是实矩阵。由式(3.23)和式(3.24)可得

$$\begin{aligned} \boldsymbol{M}_k &= (\boldsymbol{A}_k - s_2^{(k)} \boldsymbol{I}) \boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{Q}_k \boldsymbol{A}_{k+1} \boldsymbol{Q}_k^{\mathrm{T}} \boldsymbol{Q}_k \boldsymbol{R}_k - s_2^{(k)} \boldsymbol{Q}_k \boldsymbol{R}_k = \\ &\boldsymbol{Q}_k (\boldsymbol{A}_{k+1} - s_2^{(k)} \boldsymbol{I}) \boldsymbol{R}_k = \boldsymbol{Q}_k \boldsymbol{Q}_{k+1} \boldsymbol{R}_{k+1} \boldsymbol{R}_k = \widetilde{\boldsymbol{Q}}_k \widetilde{\boldsymbol{R}}_k \end{aligned} \tag{3.29}$$

其中 $\widetilde{\boldsymbol{R}}_k = \boldsymbol{R}_{k+1} \boldsymbol{R}_k$ 是上三角矩阵。因此式(3.29)的右端是实矩阵 $\boldsymbol{M}_k$ 的 QR 分解式。从而可知，式(3.25)右端的正交矩阵 $\widetilde{\boldsymbol{Q}}_k$ 可由 $\boldsymbol{M}_k$ 作 QR 分解而获得。

根据式(3.26)、(3.29)和(3.25)，可把迭代公式(3.23)从 $\boldsymbol{A}_k$ 到 $\boldsymbol{A}_{k+2}$ 的变换过程重新整理为

$$\begin{cases} \boldsymbol{A}_1 = \boldsymbol{A}^{(n-1)} \\ \boldsymbol{M}_k = \boldsymbol{A}_k^2 - s\boldsymbol{A}_k + t\boldsymbol{I} \\ \boldsymbol{M}_k = \widetilde{\boldsymbol{Q}}_k \widetilde{\boldsymbol{R}}_k \quad (\text{对 } \boldsymbol{M}_k \text{ 作 QR 分解}) \\ \boldsymbol{A}_{k+2} = \widetilde{\boldsymbol{Q}}_k^{\mathrm{T}} \boldsymbol{A}_k \widetilde{\boldsymbol{Q}}_k \\ \qquad (k = 1, 3, 5, \cdots) \end{cases} \tag{3.30}$$

其中 s 和 t 分别由式(3.27)和(3.28)计算。迭代公式(3.30)已避免了复数运算(即使位移量 $s_1^{(k)}$ 和 $s_2^{(k)}$ 为复数)。

使用带双步位移的 QR 方法(3.30)求实矩阵 $\boldsymbol{A} \in \mathbf{R}^{n \times n}$ 全部特征值的具体算法如下：

(1) 使用矩阵的拟上三角化的算法把矩阵 $\boldsymbol{A} \in \mathbf{R}^{n \times n}$ 化为拟上三角矩阵 $\boldsymbol{A}^{(n-1)}$；给定精度水平 $\varepsilon > 0$ 和迭代最大次数 L。

(2) 记 $\boldsymbol{A}_1 = \boldsymbol{A}^{(n-1)} = [a_{ij}^{(1)}]_{n \times n}$，令 $k = 1, m = n$。

(3) 如果 $|a_{m,m-1}^{(k)}| \leqslant \varepsilon$，则得到 $\boldsymbol{A}$ 的一个特征值 $a_{mm}^{(k)}$，置 $m := m - 1$(降阶)，转(4)；否则转(5)。

(4) 如果 $m = 1$，则得到 $\boldsymbol{A}$ 的一个特征值 $a_{11}^{(k)}$，转(11)；如果 $m = 0$，则直接转(11)；如果

$m>1$，则转(3)。

(5) 求二阶子阵

$$D_k=\begin{bmatrix} a_{m-1,m-1}^{(k)} & a_{m-1,m}^{(k)} \\ a_{m,m-1}^{(k)} & a_{mm}^{(k)} \end{bmatrix}$$

的两个特征值 s_1 和 s_2，即计算二次方程

$$\lambda^2-(a_{m-1,m-1}^{(k)}+a_{mm}^{(k)})\lambda+\det \boldsymbol{D}_k=0$$

的两个根 s_1 和 s_2。

(6) 如果 $m=2$，则得到 $\boldsymbol{A}$ 的两个特征值 s_1 和 s_2，转(11)；否则转 (7)。

(7) 如果 $|a_{m-1,m-2}^{(k)}|\leqslant\varepsilon$，则得到 $\boldsymbol{A}$ 的两个特征值 s_1 和 s_2，置 $m:=m-2$(降阶)，转(4)；否则转(8)。

(8) 如果 $k=L$，则计算终止，未得到 $\boldsymbol{A}$ 的全部特征值；否则转(9)。

(9) 记 $\boldsymbol{A}_k=[a_{ij}^{(k)}]_{m\times m}(1\leqslant i,j\leqslant m)$，计算

$$s=a_{m-1,m-1}^{(k)}+a_{mm}^{(k)}$$

$$t=a_{m-1,m-1}^{(k)}a_{mm}^{(k)}-a_{m,m-1}^{(k)}a_{m-1,m}^{(k)}$$

$$\boldsymbol{M}_k=\boldsymbol{A}_k^2-s\boldsymbol{A}_k+t\boldsymbol{I} \quad (\boldsymbol{I} \text{ 是 } m \text{ 阶单位矩阵})$$

$$\boldsymbol{M}_k=\boldsymbol{Q}_k\boldsymbol{R}_k \quad (\text{对 } \boldsymbol{M}_k \text{ 作 QR 分解})$$

$$\boldsymbol{A}_{k+1}=\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{Q}_k$$

(10) 置 $k:=k+1$，转(3)。

(11) $\boldsymbol{A}$ 的全部特征值已计算完毕，停止计算。

在上述算法中，从 $\boldsymbol{M}_k$ 的 QR 分解 $\boldsymbol{M}_k=\boldsymbol{Q}_k\boldsymbol{R}_k$ 到计算 $\boldsymbol{A}_{k+1}=\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{Q}_k$，实质上是

$$\boldsymbol{H}_{n-1}\cdots\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{M}_k=\boldsymbol{R}_k$$

$$\boldsymbol{A}_{k+1}=\boldsymbol{H}_{n-1}\cdots\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}_k\boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_{n-1}$$

其中 $\boldsymbol{H}_i$ 是 Householder 矩阵，且 $\boldsymbol{Q}_k=\boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_{n-1}$。参照 3.3.1 小节的 QR 分解算法，可把 $\boldsymbol{M}_k$ 的 QR 分解与 $\boldsymbol{A}_{k+1}$ 的计算用下列算法实现：

记 $\boldsymbol{B}_1=\boldsymbol{M}_k=[b_{ij}^{(1)}]_{m\times m}$，$\boldsymbol{B}_r=[b_{ij}^{(r)}]_{m\times m}$，$\boldsymbol{C}_1=\boldsymbol{A}_k$。

对于 $r=1,2,\cdots,m-1$ 执行

(1) 若 $b_{ir}^{(r)}(i=r+1,r+2,\cdots,m)$ 全为零，则令 $\boldsymbol{B}_{r+1}=\boldsymbol{B}_r$，$\boldsymbol{C}_{r+1}=\boldsymbol{C}_r$，转(5)；否则转(2)。

(2) 计算

$$d_r=\sqrt{\sum_{i=r}^{m}(b_{ir}^{(r)})^2}$$

$$c_r=-\operatorname{sgn}(b_{rr}^{(r)})\,d_r \quad (\text{若 } b_{rr}^{(r)}=0\text{，则取 } c_r=d_r)$$

$$h_r=c_r^2-c_rb_{rr}^{(r)}$$

(3) 令 $\boldsymbol{u}_r=(0,\cdots,0,b_{rr}^{(r)}-c_r,b_{r+1,r}^{(r)},\cdots,b_{mr}^{(r)})^{\mathrm{T}}\in\mathbf{R}^m$。

(4) 计算

$$\boldsymbol{v}_r=\boldsymbol{B}_r^{\mathrm{T}}\boldsymbol{u}_r/h_r$$

$$\boldsymbol{B}_{r+1}=\boldsymbol{B}_r-\boldsymbol{u}_r\boldsymbol{v}_r^{\mathrm{T}}$$

$$\boldsymbol{p}_r=\boldsymbol{C}_r^{\mathrm{T}}\boldsymbol{u}_r/h_r$$

$$\boldsymbol{q}_r=\boldsymbol{C}_r\boldsymbol{u}_r/h_r$$

$$t_r = \boldsymbol{p}_r^{\mathrm{T}}\boldsymbol{u}_r/h_r$$

$$\boldsymbol{\omega}_r = \boldsymbol{q}_r - t_r\boldsymbol{u}_r$$

$$\boldsymbol{C}_{r-1} = \boldsymbol{C}_r - \boldsymbol{\omega}_r\boldsymbol{u}_r^{\mathrm{T}} - \boldsymbol{u}_r\boldsymbol{p}_r^{\mathrm{T}}$$

(5) 继续。

此算法执行完后，就得到 $\boldsymbol{A}_{k+1}=\boldsymbol{C}_m$，所需的乘法运算次数只是 m^2 级的。

习　题

1. 设 $\boldsymbol{A}$ 是 $n\times n$ 实矩阵，使用形式为

$$\boldsymbol{u}_k = \boldsymbol{A}^k\boldsymbol{u}_0$$

的幂法产生向量 $\boldsymbol{u}_k$ 需要多少次乘法运算？使用形式为

$$\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{u}_{i-1} \quad (i = 1,2,\cdots,k)$$

的幂法产生向量 $\boldsymbol{u}_k$ 又需要多少次乘法运算？

2. 试用幂法求下列矩阵按模最大的特征值和相应的特征向量：

(1) $\begin{bmatrix} 2 & 3 & 2 \\ 10 & 3 & 4 \\ 3 & 6 & 1 \end{bmatrix}$；　　(2) $\begin{bmatrix} 5 & 30 & -48 \\ 3 & 14 & -24 \\ 3 & 15 & -25 \end{bmatrix}$。

要求近似特征值 β_k 满足 $|\beta_k-\beta_{k-1}|/|\beta_k|\leqslant 10^{-4}$。

3. 设 $n\times n$ 实矩阵 $\boldsymbol{A}$ 具有 n 个线性无关的特征向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$，其相应的特征值 $\lambda_i(i=1,2,\cdots,n)$ 满足

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant |\lambda_n|$$

任取 n 维非零向量 $\boldsymbol{v}$ 和 $\boldsymbol{u}_0$，并且 $\boldsymbol{u}_0$ 在 $\boldsymbol{x}_1$ 方向的分量 $\alpha_1\neq 0$，$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}_1\neq\boldsymbol{0}$，试证：由迭代公式

$$\begin{cases} \boldsymbol{u}_k = A\boldsymbol{u}_{k-1} \\ \beta_k = \dfrac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_k}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{k-1}} \\ \qquad (k = 1,2,\cdots) \end{cases}$$

产生的序列 $\{\beta_k\}$ 收敛于 λ_1。

4. 试用反幂法求第 2 题(1)的矩阵按模最小的特征值和相应的特征向量，要求近似特征值 β_k^{-1} 满足 $|\beta_k^{-1}-\beta_{k-1}^{-1}|/|\beta_k^{-1}|\leqslant 10^{-4}$。

5. 已知矩阵

$$\boldsymbol{A} = \begin{bmatrix} -3 & 1 & 0 \\ 1 & -3 & -3 \\ 0 & -3 & 4 \end{bmatrix}$$

的一个特征值 $\lambda\approx 5$，试用反幂法求 λ 和相应的特征向量，要求 $|\beta_k^{-1}-\beta_{k-1}^{-1}|/|\beta_k^{-1}|\leqslant 10^{-4}$。

6. 设矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的特征值 $\lambda_i(i=1,2,\cdots,n)$ 均为实数，且满足下列关系：

$$-50 < \lambda_1 < \lambda_2 < -10 \leqslant \lambda_3 \leqslant \cdots \leqslant \lambda_n$$

试分别写出用幂法(或反幂法)求 λ_1，λ_2 和 λ_n 的迭代格式。

7. 试用 Jacobi 方法计算矩阵

$$A=\begin{bmatrix}1 & 1 & 0.5\\ 1 & 1 & 0.25\\ 0.5 & 0.25 & 2\end{bmatrix}$$

的全部特征值和相应的特征向量。(取 $\varepsilon=10^{-5}$)

8. 设

$$\boldsymbol{T}=\begin{bmatrix}\boldsymbol{P}_{3\times 3} & \boldsymbol{O}_{3\times 2}\\ \boldsymbol{O}_{2\times 3} & \boldsymbol{Q}_{2\times 2}\end{bmatrix}$$

其中 $\boldsymbol{O}_{3\times 2}$ 和 $\boldsymbol{O}_{2\times 3}$ 都是零矩阵。若 λ_i 是矩阵 $\boldsymbol{P}_{3\times 3}$ 的一个特征值，相应的特征向量为 $(a_1,a_2,a_3)^{\mathrm{T}}$；$\lambda_j$ 是矩阵 $\boldsymbol{Q}_{2\times 2}$ 的一个特征值，相应的特征向量为 $(b_1,b_2)^{\mathrm{T}}$，试证：λ_i 和 λ_j 都是矩阵 $\boldsymbol{T}$ 的特征值，相应的特征向量分别为 $(a_1,a_2,a_3,0,0)^{\mathrm{T}}$ 和 $(0,0,0,b_1,b_2)^{\mathrm{T}}$。

9. 设有向量 $\boldsymbol{s}=(-2,1,2)^{\mathrm{T}}$ 和向量 $\boldsymbol{r}=(4,3,0)^{\mathrm{T}}$，试求一个 Householder 矩阵 $\boldsymbol{H}$，使 $\boldsymbol{u}=\boldsymbol{H}\boldsymbol{s}$ 成为长度与 $\boldsymbol{s}$ 相等、方向与 $\boldsymbol{r}$ 同向的向量。

10. 利用 Householder 矩阵对矩阵 $\boldsymbol{A}$ 作正交相似变换，把矩阵 $\boldsymbol{A}$ 化为三对角矩阵，其中

$$A=\begin{bmatrix}1 & 2 & 1 & 2\\ 2 & 1 & 2 & -1\\ 1 & 2 & 0 & 3\\ 2 & -1 & 3 & 1\end{bmatrix}$$

11. 利用 Householder 矩阵对矩阵

$$A=\begin{bmatrix}2 & 1 & 3 & 4\\ 1 & -1 & 2 & 1\\ -1 & 2 & 1 & 2\\ 1 & 0 & -1 & 3\end{bmatrix}$$

作正交相似变换，把 $\boldsymbol{A}$ 化为拟上三角矩阵。

12. 用基本 QR 方法计算矩阵

$$A=\begin{bmatrix}3 & 1 & 0\\ 1 & 2 & 1\\ 0 & 1 & 1\end{bmatrix}$$

的特征值。

第4章 非线性方程与非线性方程组的迭代解法

4.1 非线性方程的迭代解法

设有非线性方程

$$f(x)=0 \tag{4.1}$$

其中 $f(x)$是一元非线性函数。若常数 s 使 $f(s)=0$,则称 s 是方程(4.1)的根,又称 s 是函数 $f(x)$的零点。若 $f(x)$能分解为

$$f(x)=(x-s)^m\varphi(x)$$

其中 m 是正整数,$\varphi(s)\neq 0$,则称 s 是方程(4.1)的 m 重根和 $f(x)$的 m 重零点。当 $m=1$ 时,s 称为方程(4.1)的单根和 $f(x)$的单零点。

只有很少类型的非线性方程能解出根的解析表达式。对于大多数非线性方程,只能用数值方法求出它的根的近似值。本章将要介绍几种常用的有效的数值求根方法,它们都属于迭代法,因而还要讨论这些方法的收敛性和收敛速度。

4.1.1 对分法

用 $C[a,b]$和 $C(a,b)$分别表示在闭区间$[a,b]$和开区间(a,b)上连续的函数集合。

设 $f(x)\in C[a,b]$且 $f(a)f(b)<0$,根据连续函数的介值定理,在区间(a,b)内至少有一实数 s,使 $f(s)=0$。现假定在(a,b)内只有一个实数 s 使 $f(s)=0$,并要把 s 求出来。用对分法求这个 s 的算法如下:

令 $a_0=a,b_0=b$。

对于 $k=0,1,\cdots,M$ 执行

(1) 计算 $x_k=\dfrac{a_k+b_k}{2}$。

(2) 若 $b_k-a_k\leqslant\epsilon$ 或$|f(x_k)|\leqslant\eta$,则停止计算,取 $s\approx x_k$;否则转(3)。

(3) 若 $f(a_k)f(x_k)<0$,则令 $a_{k+1}=a_k,b_{k+1}=x_k$;若 $f(a_k)f(x_k)>0$,则令 $a_{k+1}=x_k,b_{k+1}=b_k$。

(4) 若 $k=M$,则输出 M 次迭代不成功的信息;否则继续。

上述算法中的正数 ϵ 和 η 是预先给定的精度水平。由于对任一个 k 都有

$$s\in(a_k,b_k),\quad x_k=\frac{a_k+b_k}{2}$$

所以

$$|x_k-s|<\frac{b_k-a_k}{2}=\frac{b-a}{2^{k+1}}$$

$$\lim_{k\to\infty}x_k=s$$

可见，只要函数 $f(x)\in C[a,b]$，对分法所产生的序列$\{x_k\}$必收敛于方程(4.1)在(a,b)内的根 s，收敛速度与公比为$\frac{1}{2}$的等比数列的收敛速度相同。若要求 x_k 近似 s 的绝对误差限为 ε_0，则在上述算法中取 $\varepsilon=2\varepsilon_0$，并采用 $b_k-a_k\leqslant\varepsilon$ 作为结束迭代的条件。

对分法只能求方程(4.1)的实数根，而且只能求单根和奇数重根，不能求偶数重根和复数根。

4.1.2 简单迭代法及其收敛性

设方程(4.1)有根 s，把方程(4.1)化为等价方程

$$x=\varphi(x) \tag{4.2}$$

因而有 $s=\varphi(s)$。选定 s 的初始近似值 x_0，用递推公式

$$x_{k+1}=\varphi(x_k)\quad(k=0,1,\cdots) \tag{4.3}$$

产生序列$\{x_k\}$，在一定条件下，序列$\{x_k\}$收敛于 s。在$\{x_k\}$收敛的情况下，当 k 足够大时就可取 x_k 作为方程(4.1)的近似根。迭代公式(4.3)被称为求解方程(4.1)的简单迭代法，其中 $\varphi(x)$称为迭代函数。因 $s=\varphi(s)$，故 s 是迭代函数 $\varphi(x)$的不动点。简单迭代法(4.3)又称为不动点迭代法。

把方程(4.1)化为等价方程(4.2)可以有多种方案。例如方程

$$x^3-x-1=0$$

可化为以下三种等价方程

$$x=x^3-1,\quad x=\frac{x+1}{x^2},\quad x=\sqrt[3]{x+1}$$

由此可构成多种简单迭代法，它们的迭代函数各不相同。为求同一个根，它们所产生的序列$\{x_k\}$，有的可能收敛，有的可能不收敛；有的会收敛得快，有的会收敛得慢。这一切都取决于迭代函数 $\varphi(x)$在有根区间内的性态。

定理 4.1 设函数 $\varphi(x)\in C[a,b]$，在(a,b)内可导，且满足两个条件：

(1) 当 $x\in[a,b]$时，$\varphi(x)\in[a,b]$；

(2) 当 $x\in(a,b)$时，$|\varphi'(x)|\leqslant L<1$，其中 L 为一常数。

则有如下结论：

(1) 方程(4.2)在区间$[a,b]$上有唯一的根 s；

(2) 对任取的 $x_0\in[a,b]$，简单迭代法(4.3)产生的序列$\{x_k\}\subset[a,b]$且收敛于 s；

(3) 成立误差估计式

$$|s-x_k|\leqslant\frac{L^k}{1-L}|x_1-x_0| \tag{4.4}$$

$$|s-x_k|\leqslant\frac{L}{1-L}|x_k-x_{k-1}| \tag{4.5}$$

证 (1) 令 $F(x)=x-\varphi(x)$，则 $F(x)\in C[a,b]$，并由条件(1)可知

$$F(a)=a-\varphi(a)\leqslant 0,\quad F(b)=b-\varphi(b)\geqslant 0$$

若上面两个不等式中有一个等号成立，则方程(4.2)有根 $s=a$ 或$s=b$；若两个都是严格不等式，则根据连续函数的介值定理，必存在 $s\in(a,b)$，使 $F(s)=s-\varphi(s)=0$，即方程(4.2)有根 $s\in(a,b)$。今设有两个不同的 $s_1,s_2\in[a,b]$使 $s_1=\varphi(s_1)$，$s_2=\varphi(s_2)$，则由微分中值定理以及条件(2)，有

$$|s_1-s_2|=|\varphi(s_1)-\varphi(s_2)|=|\varphi'(\xi)||s_1-s_2|\leqslant$$

$$L \mid s_1 - s_2 \mid < \mid s_1 - s_2 \mid$$

其中 ξ 在 s_1 与 s_2 之间，因而 $\xi \in (a,b)$。上式出现的矛盾证实 $s_1 = s_2$。

(2) 因 $x_0 \in [a,b]$，由条件(1)可知，$\{x_k\} \subset [a,b]$。又由条件(2)，得

$$\mid x_k - s \mid = \mid \varphi(x_{k-1}) - \varphi(s) \mid = \mid \varphi'(\xi_k) \mid \mid x_{k-1} - s \mid \leqslant$$
$$L \mid x_{k-1} - s \mid \leqslant \cdots \leqslant L^k \mid x_0 - s \mid$$

其中 ξ_k 在 x_{k-1} 与 s 之间，因而 $\xi_k \in (a,b)$。因 $0 \leqslant L < 1$，故有 $\lim\limits_{k\to\infty} x_k = s$。

(3) 设 $m > k$，则有

$$x_m - x_k = \sum_{i=k}^{m-1} (x_{i-1} - x_i)$$

而

$$\mid x_{i+1} - x_i \mid = \mid \varphi(x_i) - \varphi(x_{i-1}) \mid \leqslant$$
$$L \mid x_i - x_{i-1} \mid \leqslant \cdots \leqslant L^i \mid x_1 - x_0 \mid$$

于是有

$$\mid x_m - x_k \mid \leqslant \sum_{i=k}^{m-1} \mid x_{i-1} - x_i \mid \leqslant \sum_{i=k}^{m-1} L^i \mid x_1 - x_0 \mid =$$
$$L^k \frac{1 - L^{m-k}}{1 - L} \mid x_1 - x_0 \mid$$

令 $m \to \infty$，由于 $0 \leqslant L < 1$，故由上式得到式(4.4)。

又由

$$\mid x_{i+1} - x_i \mid \leqslant L \mid x_i - x_{i-1} \mid \leqslant \cdots \leqslant L^{i-k+1} \mid x_k - x_{k-1} \mid$$

得到

$$\mid x_m - x_k \mid \leqslant \sum_{i=k}^{m-1} \mid x_{i+1} - x_i \mid \leqslant \sum_{i=k}^{m-1} L^{i-k+1} \mid x_k - x_{k-1} \mid =$$
$$L \frac{1 - L^{m-k}}{1 - L} \mid x_k - x_{k-1} \mid$$

令 $m \to \infty$，由上式得到式(4.5)。

证毕。

实际计算时，可预先给定精度水平 $\eta > 0$，当式(4.3)产生的 x_k 满足

$$\frac{\mid x_k - x_{k-1} \mid}{\mid x_k \mid} \leqslant \eta$$

时，迭代结束，用当前的 x_k 作为方程(4.2)的近似根。也可以利用式(4.4)，根据预先给出的近似根的绝对误差限 ε，求出满足 $|s - x_k| \leqslant \varepsilon$ 的最低迭代次数 N。由

$$\frac{L^k}{1 - L} \mid x_1 - x_0 \mid \leqslant \varepsilon$$

解得

$$k \geqslant \frac{\ln \dfrac{\varepsilon(1 - L)}{\mid x_1 - x_0 \mid}}{\ln L} = d$$

因此，$N = [d] + 1$，其中 $[d]$ 表示不大于 d 的最大整数。

定理 4.1 指定了一个固定的区间 $[a,b]$，在此区间内任取一点 x_0 作为初值，迭代都收敛。

这种形式的收敛定理称为大范围收敛性定理。但当条件不够充分时，预先指定一个区间常常是不可能的。若能设法使初值 x_0 充分接近根 s，则仍然可以希望迭代收敛。描述成“存在根 s 的一个邻域，当初值 x_0 在此邻域内任取时，迭代都能收敛”这种形式的定理称为局部收敛性定理。下面就是简单迭代法的局部收敛性定理。

定理 4.2 设 $s=\varphi(s)$，$\varphi'(x)$在包含 s 的某个开区间内连续。如果$|\varphi'(s)|<1$，则存在 $\delta>0$，当 $x_0\in[s-\delta,s+\delta]$时，由简单迭代法(4.3)产生的序列$\{x_k\}\subset[s-\delta,s+\delta]$且收敛于 s。

证 总可取一常数 L，使$|\varphi'(s)|<L<1$。由 $\varphi'(x)$的连续性，必存在 $\delta>0$，当 $x\in(s-\delta,s+\delta)$时，$|\varphi'(x)|\leqslant L$；又根据微分中值定理，当 $x\in[s-\delta,s+\delta]$时，有

$$|\varphi(x)-s|=|\varphi(x)-\varphi(s)|=|\varphi'(\xi)||x-s|$$

其中 ξ 在 x 与 s 之间，因而有

$$|\varphi(x)-s|\leqslant L|x-s|<\delta$$

即 $\varphi(x)\in[s-\delta,s+\delta]$。根据定理 4.1，对任取的 $x_0\in[s-\delta,s+\delta]$，由简单迭代法(4.3)产生的序列$\{x_k\}\subset[s-\delta,s+\delta]$且收敛于 s。

证毕。

定理 4.2 并没有指出 δ 的值是多少，它仅仅证实 δ 的存在。因而在满足定理的条件时只要 x_0 足够接近 s，由迭代公式(4.3)产生的序列$\{x_k\}$就收敛于 s。

例 1 用简单迭代法求方程 $x-\ln x=2$ 在区间$(2,\infty)$内的根，要求$\dfrac{|x_k-x_{k-1}|}{|x_k|}\leqslant 10^{-8}$。

解 记 $f(x)=x-\ln x-2$，因 $f(2)<0$，$f(4)>0$，故方程在区间(2,4)内有根。又因

$$f'(x)=1-\frac{1}{x}>0,\quad x\in(2,\infty)$$

故 $f(x)$在区间$(2,\infty)$上单调增大，因而方程在区间$(2,\infty)$内仅有一个根 s，且 $s\in(2,4)$。

今把所给方程化为等价方程

$$x=2+\ln x$$

因而得到迭代函数 $\varphi(x)=2+\ln x$，且 $\varphi(x)$满足

$$\varphi(x)\in[2+\ln 2,2+\ln 4]\subset[2,4],\quad x\in[2,4]$$

$$|\varphi'(x)|=\left|\frac{1}{x}\right|<0.5<1,\quad x\in(2,4)$$

根据定理 4.1，对任取的 $x_0\in[2,4]$，由迭代公式

$$x_{k+1}=2+\ln x_k\quad(k=0,1,\cdots)$$

产生的序列$\{x_k\}$必收敛于 s。现在取 $x_0=3$，迭代结果见表 4-1。

表 4-1 例 1 计算结果

k	x_k	k	x_k
0	3.000 000 000	8	3.146 177 452
1	3.098 612 289	9	3.146 188 209
2	3.130 954 363	10	3.146 191 628
3	3.141 337 866	11	3.146 192 714
4	3.144 648 781	12	3.146 193 060
5	3.145 702 209	13	3.146 193 170
6	3.146 037 143	14	3.146 193 205
7	3.146 143 611	15	3.146 193 216

由于

$$\frac{|x_{15}-x_{14}|}{|x_{15}|}=0.35\times10^{-8}<10^{-8}$$

所以，取 $s\approx x_{15}=3.146\ 193\ 216$ 为方程的近似根。

4.1.3　简单迭代法的收敛速度

定义　设序列 $\{x_k\}$ 收敛于 s，并且 $e_k=s-x_k\neq0(k=0,1,\cdots)$，如果存在常数 $r\geqslant1$ 和常数 $c>0$，使得极限

$$\lim_{k\to\infty}\frac{|e_{k+1}|}{|e_k|^r}=c$$

成立，或者使得当 $k\geqslant K$（某个正整数）时，

$$\frac{|e_{k+1}|}{|e_k|^r}\leqslant c$$

成立，则称序列 $\{x_k\}$ 收敛于 s 具有 r 阶收敛速度，简称 $\{x_k\}$ 是 r 阶收敛的。常数 c 称为渐近收敛常数，也称为收敛因子。

显然，r 的大小反映了序列 $\{x_k\}$ 收敛的快慢程度，r 越大收敛越快。$r=1$ 时，又称序列 $\{x_k\}$ 是线性收敛的，此时必有 $0<c\leqslant1$；$r=2$ 时，又称序列 $\{x_k\}$ 是平方收敛的；对于 $r>1$ 的情况，序列 $\{x_k\}$ 统称为是超线性收敛的。

定理 4.3　设函数 $\varphi(x)\in C[a,b]$，$\varphi'(x)\in C(a,b)$，且满足如下条件：

(1) 当 $x\in[a,b]$ 时，$\varphi(x)\in[a,b]$；

(2) 当 $x\in(a,b)$ 时，$\varphi'(x)\neq0$，$|\varphi'(x)|\leqslant L<1$，其中 L 为一常数。

则对任取的 $x_0\in[a,b]$，由简单迭代法(4.3)产生的序列 $\{x_k\}$ 收敛于方程(4.2)在 $[a,b]$ 内的唯一的根 s，并且当 $x_0\neq s$ 时 $\{x_k\}$ 是线性收敛的。

证　因定理所给条件包含了定理 4.1 的两个条件，所以对任取的 $x_0\in[a,b]$，由迭代法(4.3)产生的序列 $\{x_k\}$ 收敛于方程(4.2)在 $[a,b]$ 内的唯一的根 s。下面证明当 $x_0\neq s$ 时，$\{x_k\}$ 是线性收敛的。

由于当 $x\in(a,b)$ 时 $\varphi'(x)\neq0$，所以，只要 $x_0\neq s$ 就必有 $x_k\neq s(k=1,2,\cdots)$。由 Taylor 公式，有

$$\varphi(x_k)=\varphi(s)+\varphi'(s+\theta(x_k-s))(x_k-s),\quad 0<\theta<1$$

记 $e_k=s-x_k$，由上式得

$$e_{k+1}=\varphi'(s-\theta e_k)e_k$$

因 $\lim\limits_{k\to\infty}e_k=0$ 和 $\varphi'(x)$ 的连续性，故得

$$\lim_{k\to\infty}\frac{|e_{k+1}|}{|e_k|}=\lim_{k\to\infty}|\varphi'(s-\theta e_k)|=|\varphi'(s)|$$

因为 $|\varphi'(s)|>0$，所以由上式可知，序列 $\{x_k\}$ 是线性收敛的。

证毕。

由定理 4.3 看到，迭代函数 $\varphi(x)$ 在方程的根 s 处的一阶导数不等于零，相应的简单迭代法只是线性收敛的，也就是说，只有一阶收敛速度。要想获得高阶收敛速度，迭代函数 $\varphi(x)$ 就要满足更多的条件。

定理 4.4 设 $s=\varphi(s)$，$\varphi^{(m)}(x)$在包含 s 的某个开区间内连续($m\geqslant 2$)。如果

$$\varphi^{(i)}(s)=0 \quad (i=1,2,\cdots,m-1)$$
$$\varphi^{(m)}(s)\neq 0$$

则存在 $\delta>0$，当 $x_0\in[s-\delta,s+\delta]$但 $x_0\neq s$ 时，由简单迭代法(4.3)产生的序列$\{x_k\}$以 m 阶收敛速度收敛于 s。

证 由 $\varphi^{(m)}(s)\neq 0$ 以及 $\varphi^{(m)}(x)$的连续性可知，必存在 $\delta_1>0$，使当 $x\in[s-\delta_1,s+\delta_1]$时，$\varphi^{(m)}(x)\neq 0$。又根据定理 4.2，必存在 $\delta>0$(取 $\delta\leqslant\delta_1$)，当 $x_0\in[s-\delta,s+\delta]$时，由迭代公式(4.3)产生的序列$\{x_k\}$收敛于 s，并且有 $x_k\in[s-\delta,s+\delta]$($k=1,2,\cdots$)。

由 Taylor 公式，得

$$\varphi(x_k)=\varphi(s)+\varphi'(s)(x_k-s)+\cdots+\frac{\varphi^{(m-1)}(s)}{(m-1)!}(x_k-s)^{m-1}+$$
$$\frac{\varphi^{(m)}(s+\theta(x_k-s))}{m!}(x_k-s)^m,\quad 0<\theta<1$$

利用定理的条件，得

$$e_{k+1}=\varphi(s)-\varphi(x_k)=(-1)^{m+1}\frac{\varphi^{(m)}(s-\theta e_k)}{m!}e_k^m$$

因 $s-\theta e_k\in[s-\delta,s+\delta]$，故

$$\varphi^{(m)}(s-\theta e_k)\neq 0 \quad (k=0,1,\cdots)$$

因而只要 $e_0=s-x_0\neq 0$ 就有 $e_k=s-x_k\neq 0$($k=1,2,\cdots$)。于是有

$$\frac{|e_{k+1}|}{|e_k|^m}=\frac{|\varphi^{(m)}(s-\theta e_k)|}{m!}$$

由于$\lim\limits_{k\to\infty}e_k=0$ 以及 $\varphi^{(m)}(x)$的连续性，故得

$$\lim_{k\to\infty}\frac{|e_{k+1}|}{|e_k|^m}=\frac{|\varphi^{(m)}(s)|}{m!}$$

因$|\varphi^{(m)}(s)|>0$，故由上式可知，序列$\{x_k\}$具有 m 阶收敛速度。

证毕。

例如，为了用简单迭代法求方程

$$x^3-x-1=0$$

的正实根 s，第一种方案是把方程改写成

$$x=\sqrt[3]{x+1}$$

得迭代函数 $\varphi(x)=\sqrt[3]{x+1}$。因 $\varphi'(x)=\dfrac{1}{3(x+1)^{2/3}}$在 s 的某邻域内连续，且$|\varphi'(s)|<1$，但 $\varphi'(s)\neq 0$，所以，当 x_0 充分接近 s 时，迭代公式

$$x_{k+1}=\sqrt[3]{x_k+1} \quad (k=0,1,\cdots)$$

产生的序列$\{x_k\}$收敛于 s，但只有线性收敛速度。第二种方案是把原方程改写成

$$x=\frac{2x^3+1}{3x^2-1}$$

得迭代函数

$$\varphi(x)=\frac{2x^3+1}{3x^2-1}$$

这时 $\varphi''(x)$在 s 的某邻域内连续，并且由

$$\varphi'(x) = \frac{6x(x^3 - x - 1)}{(3x^2 - 1)^2}$$

可知 $\varphi'(s)=0$，所以，只要 x_0 充分接近 s，由迭代公式

$$x_{k+1} = \frac{2x_k^3 + 1}{3x_k^2 - 1} \quad (k = 0,1,\cdots)$$

产生的序列$\{x_k\}$以至少是二阶的收敛速度收敛于 s。

4.1.4 Steffensen 迭代法

从 4.1.3 小节的讨论知道，当迭代函数 $\varphi(x)$在方程(4.2)的根 s 处的导数不为零时，简单迭代法(4.3)即使收敛于 s，也只是线性收敛的。现在来研究，在式(4.3)产生的序列$\{x_k\}$线性收敛于 s 的情况下能不能加速收敛。

设已由式(4.3)计算出 x_k, x_{k+1}, x_{k+2}三个迭代值，由微分中值定理可得

$$x_{k+1} - s = \varphi(x_k) - \varphi(s) = \varphi'(\xi_k)(x_k - s)$$

$$x_{k+2} - s = \varphi(x_{k+1}) - \varphi(s) = \varphi'(\xi_{k+1})(x_{k+1} - s)$$

其中 ξ_k 在 x_k 与 s 之间，ξ_{k+1}在 x_{k+1}与 s 之间。假定当 x 在 s 的附近变化时 $\varphi'(x)$变化不大，则有

$$\frac{x_{k+1} - s}{x_k - s} \approx \frac{x_{k+2} - s}{x_{k+1} - s}$$

由此解出

$$s \approx x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k}$$

上式右端的值可能会比 x_{k+1}, x_{k+2}更接近 s。因此，可把这个右端值作为继 x_k 之后的一个新的 x_{k+1}，而原 x_{k+1}和原 x_{k+2}只起中间值的作用，并把它们分别记为 y_k 和 z_k。于是得到以下的迭代公式

$$\begin{cases} y_k = \varphi(x_k), \quad z_k = \varphi(y_k) \\ x_{k+1} = x_k - \dfrac{(y_k - x_k)^2}{z_k - 2y_k + x_k} \\ \qquad (k = 0,1,\cdots) \end{cases} \tag{4.6}$$

迭代公式(4.6)称为 Steffensen(斯蒂芬森)迭代法。

定理 4.5 设 $s=\varphi(s)$，$\varphi(x)$在包含 s 的某个开区间内具有二阶连续导数，并且 $\varphi'(s)\neq 1$，则存在 $\delta>0$，当 $x_0\in[s-\delta, s+\delta]$但 $x_0\neq s$ 时，由 Setffensen 迭代法(4.6)产生的序列$\{x_k\}$至少以二阶收敛速度收敛于 s。

证 构造函数 $\psi(x)$：

当 $x=s$ 时 $\quad \psi(x)=s$

当 $x\neq s$ 时 $\quad \psi(x)=x-\dfrac{[\varphi(x)-x]^2}{\varphi(\varphi(x))-2\varphi(x)+x}=\dfrac{x\varphi(\varphi(x))-[\varphi(x)]^2}{\varphi(\varphi(x))-2\varphi(x)+x}$

于是，方程 $x=\psi(x)$与方程(4.2)有共同的根 s，并且，迭代公式(4.6)就是以 $\psi(x)$为迭代函数的简单迭代法：

$$x_{k+1} = \psi(x_k) \quad (k = 0,1,\cdots)$$

利用 L'Hôpital(罗彼塔)法则可得

$$\lim_{x\to s}\psi(x)=s,\quad \lim_{x\to s}\psi'(x)=0$$

$$\lim_{x\to s}\psi''(x)=\frac{[3\varphi'(s)-4]\varphi''(s)}{3[\varphi'(s)-1]}$$

又由于 $\varphi'(s)\neq 1$ 以及 $\varphi''(x)$ 的连续性可知,函数 $\psi(x)$ 在包含 s 的某个开区间内具有二阶连续导数,并且 $\psi'(s)=0$。根据定理 4.4,必存在 $\delta>0$,当 $x_0\in[s-\delta,s+\delta]$ 但 $x_0\neq s$ 时,Steffensen 迭代法(4.6)产生的序列 $\{x_k\}$ 至少以二阶收敛速度收敛于 s。

证毕。

由于当迭代法(4.3)线性收敛时,由式(4.3)的迭代函数构造出来的迭代法(4.6)能平方收敛,所以又称(4.6)为 Steffensen 加速收敛方法。实际上,只要 $\varphi(x)$ 满足定理 4.5 的条件,则无论迭代法(4.3)是否收敛于 s,迭代法(4.6)都能以不低于二阶的收敛速度收敛于 s。但是,如果迭代法(4.3)已经具有 $p(p\geqslant 2)$ 阶收敛速度,则使用迭代法(4.6)对提高收敛速度作用不大。

例 2 试分别采用 $\varphi(x)=2+\ln x$ 和 $\varphi(x)=e^{x-2}$ 的 Steffensen 迭代法求方程 $x-\ln x=2$ 在区间 $(2,\infty)$ 内的根 s,精度要求与例 1 相同。

解 对于 $\varphi(x)=2+\ln x$, $\varphi'(s)=\frac{1}{s}\neq 1$,对于 $\varphi(x)=e^{x-2}$, $\varphi'(s)=e^{s-2}\neq 1$,故当 x_0 足够接近 s 时,迭代公式

$$\begin{cases} y_k=2+\ln x_k \\ z_k=2+\ln y_k \\ x_{k+1}=x_k-\dfrac{(y_k-x_k)^2}{z_k-2y_k+x_k} \\ \qquad (k=0,1,\cdots) \end{cases}$$

和

$$\begin{cases} y_k=e^{x_k-2} \\ z_k=e^{y_k-2} \\ x_{k+1}=x_k-\dfrac{(y_k-x_k)^2}{z_k-2y_k+x_k} \\ \qquad (k=0,1,\cdots) \end{cases}$$

都能平方收敛于 s。现取 $x_0=3$,前者的迭代过程见表 4-2, $s\approx x_4=3.146\ 193\ 220$ 已满足精度要求;后者的迭代过程见表 4-3。$s\approx x_5=3.146\ 193\ 262$ 已满足精度要求。

表 4-2 例 2 计算结果 1

k	x_k
0	3.000 000 000
1	3.146 738 373
2	3.146 245 819
3	3.146 193 220
4	3.146 193 220

表 4-3 例 2 计算结果 2

k	x_k
0	3.000 000 000
1	3.205 791 857
2	3.153 859 280
3	3.146 327 554
4	3.146 193 262
5	3.146 193 262

值得注意的是,迭代公式

$$\begin{cases} x_0 = 3 \\ x_{k+1} = e^{x_k - 2} \quad (k = 0,1,\cdots) \end{cases}$$

并不收敛于方程 $x-\ln x=2$ 在区间$(2,\infty)$内的根。

从 4.1.2 小节到 4.1.4 小节都只就方程(4.1)的实数根来讨论。但类似的结果完全可以推广到求方程的复数根中去。也就是说，简单迭代法(4.3)和 Steffensen 迭代法(4.6)可用于求方程的复数根，这时 x_0 是复平面上有根区域内的一点，$\{x_k\}$是复数序列。

4.1.5 Newton 法

用简单迭代法求方程(4.1)的根 s，十分重要的问题是构造迭代函数 $\varphi(x)$。为了使收敛速度的阶高一些，应尽可能使 $\varphi(x)$在$x=s$处有更多阶导数等于零。

现在令 $\varphi(x)=x+h(x)f(x)$，$h(x)$为待定函数，但 $h(s)\neq 0$，则方程(4.1)与方程

$$x = x + h(x)f(x)$$

有共同的根 s。现用条件 $\varphi'(s)=0$ 确定 $h(x)$。由

$$\varphi'(s) = 1 + h'(s)f(s) + h(s)f'(s) = 1 + h(s)f'(s) = 0$$

知道，$h(x)$必须满足 $h(s)=\dfrac{-1}{f'(s)}$。显然，取 $h(x)=-\dfrac{1}{f'(x)}$就具备这个条件，并且也满足 $h(s)\neq 0$。于是，$\varphi(x)$被确定为

$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$

它满足 $\varphi'(s)=0$。由此得出下面的特殊的简单迭代法

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (k = 0,1,\cdots) \tag{4.7}$$

式(4.7)所表示的迭代法称为 Newton(牛顿)法。

Newton 法可求方程(4.1)的实数根和复数根。当求实数根时，Newton 法有明显的几何意义。当获得 x_k 之后，过曲线 $y=f(x)$上的点$(x_k, f(x_k))$作该曲线的切线，此切线与 x 轴相交的交点横坐标就是 Newton 法迭代序列的第 $k+1$ 个元素 x_{k+1}，见图4-1。事实上，该切线方程为

$$y = f'(x_k)(x - x_k) + f(x_k)$$

令 $y=0$，可得

$$x = x_k - \frac{f(x_k)}{f'(x_k)} = x_{k+1}$$

因此 Newton 法(4.7)又称为切线法。

定理 4.6　设 s 是方程(4.1)的根，在包含 s 的某个开区间内 $f''(x)$连续且 $f'(x)\neq 0$，则存在 $\delta>0$，当 $x_0\in[s-\delta, s+\delta]$时，由 Newton 法(4.7)产生的序列$\{x_k\}$收敛于 s；若 $f''(s)\neq 0$ 且 $x_0\neq s$，则序列$\{x_k\}$是平方收敛的。

证　Newton 法(4.7)的迭代函数 $\varphi(x)$为

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \quad \varphi'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$$

因 $f''(x)$连续且 $f'(x)\neq 0$，故 $\varphi'(x)$在包含 s 的某个开区间内连续，并且有

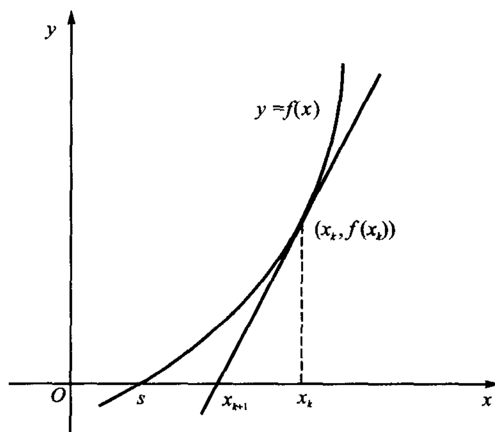


图 4-1 切线法

$$\varphi(s) = s, \quad \varphi'(s) = 0$$

根据定理 4.2，必存在 $\delta>0$，当 $x_0\in[s-\delta,s+\delta]$ 时，由迭代法(4.7)产生的序列 $\{x_k\}\subset[s-\delta,s+\delta]$，且收敛于 s。

由 Taylor 公式以及 $f(s)=0$，有

$$f(s) = f(x_k) + f'(x_k)(s - x_k) + \frac{f''(\xi_k)}{2!}(s - x_k)^2 = 0$$

$$x_{k+1} - s = x_k - \frac{f(x_k)}{f'(x_k)} - s = \frac{f''(\xi_k)}{2f'(x_k)}(s - x_k)^2$$

其中 ξ_k 在 x_k 与 s 之间。因 $f''(s)\neq 0$ 以及 $f''(x)$ 的连续性，故可设当 $x\in[s-\delta,s+\delta]$ 时 $f''(x)\neq 0$。因而当 $x_0\in[s-\delta,s+\delta]$ 且 $x_0\neq s$ 时，$x_k\neq s(k=1,2,\cdots)$。于是有

$$\lim_{k\to\infty}\frac{|x_{k+1}-s|}{|x_k-s|^2} = \left|\frac{f''(s)}{2f'(s)}\right| > 0$$

故序列 $\{x_k\}$ 是平方收敛的。

证毕。

定理 4.6 是 Newton 法的局部收敛性定理。如果 $f(x)$ 只满足此定理的条件，则在一般情况下，初值 x_0 要比较靠近所要求的根 s，Newton 法才能收敛，也就是定理结论中的 δ 要比较小。下面给出一个大范围收敛性定理。

定理 4.7 设函数 $f(x)$ 在区间 $[a,b]$ 上存在二阶连续导数，且满足条件：

(1) $f(a)f(b)<0$；

(2) $f''(x)$ 在区间 $[a,b]$ 上不变号；

(3) 当 $x\in[a,b]$ 时，$f'(x)\neq 0$；

(4) $x_0\in[a,b]$，$f(x_0)f''(x_0)>0$。

则由 Newton 法(4.7)产生的序列 $\{x_k\}$ 单调收敛于方程(4.1)在 $[a,b]$ 内唯一的根 s，并且至少是平方收敛的。

证 因 $f(x)$ 在 $[a,b]$ 上连续，由条件(1)可知，方程(4.1)在 (a,b) 内有根 s。又由条件(3)知道 $f'(x)$ 在 $[a,b]$ 上恒正或恒负，所以 $f(x)$ 在 $[a,b]$ 上严格单调，因而方程(4.1)在 (a,b) 内的

根 s 是唯一的。

条件(1),(2)共有四种情形：

(1) $f(a)<0, f(b)>0$,当 $x\in[a,b]$时 $f''(x)\leqslant 0$；

(2) $f(a)<0, f(b)>0$,当 $x\in[a,b]$时 $f''(x)\geqslant 0$；

(3) $f(a)>0, f(b)<0$,当 $x\in[a,b]$时 $f''(x)\leqslant 0$；

(4) $f(a)>0, f(b)<0$,当 $x\in[a,b]$时 $f''(x)\geqslant 0$。

现只就情形(1)进行定理的证明(见图 4-2),其余三种情形的证明方法是类似的。

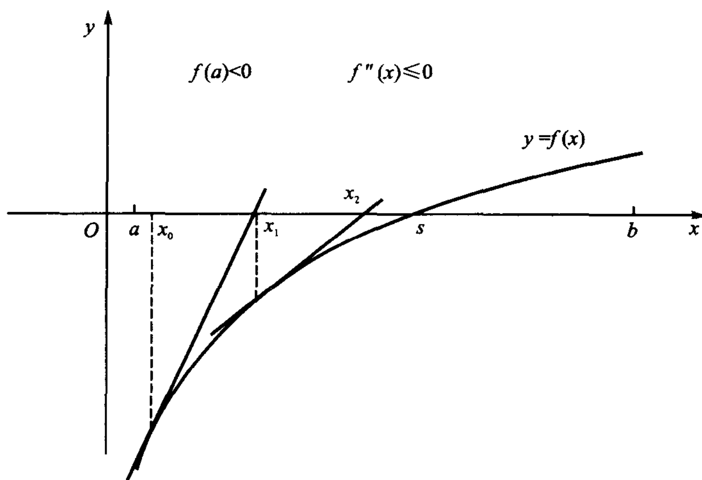


图 4-2　定理 4.7 的情形(1)

由微分中值定理,存在 $\xi\in(a,b)$使

$$f'(\xi)=\frac{f(b)-f(a)}{b-a}>0$$

因而 $f'(x)$在$[a,b]$上恒正。

由 $x_0\in[a,b], f(x_0)f''(x_0)>0$ 可知 $f(x_0)<0, x_0<s$。由式(4.7),有

$$x_1=x_0-\frac{f(x_0)}{f'(x_0)}>x_0$$

另一方面,由微分中值定理,有

$$x_1=x_0-\frac{f(s)+f'(\xi_0)(x_0-s)}{f'(x_0)}=x_0+\frac{f'(\xi_0)}{f'(x_0)}(s-x_0),\quad \xi_0\in(x_0,s)$$

因 $f''(x)\leqslant 0, x\in[a,b]$,故 $f'(x)$在$[a,b]$上单调减小,因而 $f'(x_0)\geqslant f'(\xi_0)>0$,从而有

$$0<\frac{f'(\xi_0)}{f'(x_0)}\leqslant 1$$

$$x_1\leqslant x_0+(s-x_0)=s$$

一般地,设 $a<x_k\leqslant s$,则必有 $f(x_k)\leqslant 0$,因 $f'(x)$在$[a,b]$上恒正,所以有

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}\geqslant x_k$$

另一方面,由微分中值定理,得

$$x_{k+1} = x_k - \frac{f(s) + f'(\xi_k)(x_k - s)}{f'(x_k)} = x_k + \frac{f'(\xi_k)}{f'(x_k)}(s - x_k), \quad \xi_k \in (x_k, s)$$

由 $f'(x)$ 的性质知

$$0 < \frac{f'(\xi_k)}{f'(x_k)} \leqslant 1$$

故有

$$x_{k+1} \leqslant x_k + (s - x_k) = s$$

由此可知，式(4.7)产生的序列$\{x_k\}$单调增大，且有上界 s，因而必有极限，设为 $\tilde{s}$，$\tilde{s} \in [x_0, s]$。令 $k \to \infty$，对式(4.7)两边取极限，得

$$\tilde{s} = \tilde{s} - \frac{f(\tilde{s})}{f'(\tilde{s})}$$

因而 $f(\tilde{s})=0$。但方程(4.1)在$[a,b]$内的根 s 是唯一的，所以，$\tilde{s}=s$。另根据定理 4.6，可知序列$\{x_k\}$的收敛速度至少是平方收敛的。

证毕。

例 3 用 Newton 法求方程 $x-\ln x=2$ 在区间$(2,\infty)$内的根，要求$\frac{|x_k - x_{k-1}|}{|x_k|} < 10^{-8}$。

解 在例 1 中已分析过，此方程在区间$(2,\infty)$内只有一个根 s，而且在区间$(2,4)$内。对于此方程

$$f(x) = x - \ln x - 2, \quad f'(x) = 1 - \frac{1}{x}, \quad f''(x) = \frac{1}{x^2}$$

易知，在根 s 的某邻域内 $f''(x)$ 连续且 $f'(x) \neq 0$，因此，在靠近 s 处取初始值 x_0，Newton 法迭代公式

$$x_{k-1} = x_k - \frac{x_k - \ln x_k - 2}{1 - \frac{1}{x_k}} = \frac{x_k(1 + \ln x_k)}{x_k - 1} \quad (k = 0,1,\cdots)$$

产生的序列$\{x_k\}$将收敛于 s。今取 $x_0=3$，迭代结果见表 4-4。$s \approx x_4 = 3.146\ 193\ 221$ 已满足精度要求。

表 4-4 例 3 计算结果

k	x_k
0	3.000 000 000
1	3.147 918 433
2	3.146 193 441
3	3.146 193 221
4	3.146 193 221

对于此题，由于 $f(2)<0$，$\lim\limits_{x \to +\infty} f(x) = +\infty$，在区间$[2,\infty)$内 $f'(x)$和 $f''(x)$都恒正，因此，根据定理 4.7，对任取的 $x_0 > s$，Newton 法(4.7)产生的序列$\{x_k\}$都能单调减小地收敛于 s。

4.1.6 求方程 m 重根的 Newton 法

在 4.1.5 小节的讨论中，都是假定了在方程(4.1)的根 s 的某邻域内 $f'(x) \neq 0$，即 s 是方程(4.1)的单根。本小节要讨论方程(4.1)有多重根时 Newton 法的收敛情况。

设 s 是方程(4.1)的 m 重根($m \geqslant 2$)，$f(x)$在 s 的某邻域内有 m 阶连续导数，这时

$$f(s) = f'(s) = \cdots = f^{(m-1)}(s) = 0, \quad f^{(m)}(s) \neq 0$$

由 Taylor 公式，得

$$f(x)=\frac{f^{(m)}(\xi_1)}{m!}(x-s)^m$$

$$f'(x)=\frac{f^{(m)}(\xi_2)}{(m-1)!}(x-s)^{m-1}$$

$$f''(x)=\frac{f^{(m)}(\xi_3)}{(m-2)!}(x-s)^{m-2}$$

其中 ξ_1,ξ_2,ξ_3 都在 x 与 s 之间。由 Newton 法的迭代函数 $\varphi(x)=x-f(x)/f'(x)$，可得

$$\varphi(s)=\lim_{x\to s}\varphi(x)=\lim_{x\to s}\left[x-\frac{(x-s)f^{(m)}(\xi_1)}{mf^{(m)}(\xi_2)}\right]=s$$

$$\varphi'(s)=\lim_{x\to s}\varphi'(x)=\lim_{x\to s}\frac{f(x)f''(x)}{[f'(x)]^2}=$$

$$\lim_{x\to s}\frac{(m-1)f^{(m)}(\xi_1)f^{(m)}(\xi_3)}{m[f^{(m)}(\xi_2)]^2}=1-\frac{1}{m}$$

由此可见，方程(4.1)的 m 重根 s 仍然是其等价方程(4.2)的根。由于 $0<\varphi'(s)<1$，所以，只要 x_0 充分靠近 s，由 Newton 法(4.7)产生的序列 $\{x_k\}$ 仍收敛于 s，但是只有线性的收敛速度。

若把迭代函数修改为

$$\tilde{\varphi}(x)=x-\frac{mf(x)}{f'(x)}$$

则有

$$\tilde{\varphi}(s)=\lim_{x\to s}\tilde{\varphi}(x)=\lim_{x\to s}\left[x-\frac{(x-s)f^{(m)}(\xi_1)}{f^{(m)}(\xi_2)}\right]=s$$

$$\tilde{\varphi}'(s)=\lim_{x\to s}\tilde{\varphi}'(x)=\lim_{x\to s}\left\{1-m+\frac{mf(x)f''(x)}{[f'(x)]^2}\right\}=$$

$$1-m+m\left(1-\frac{1}{m}\right)=0$$

这两个等式说明，当 s 是方程(4.1)的 m 重根时，变形的Newton 法：

$$x_{k+1}=x_k-\frac{mf(x_k)}{f'(x_k)}\quad(k=0,1,\cdots)\tag{4.8}$$

不仅可以收敛于 s，而且还具有二阶的收敛速度。

在重根的情况下，一般重数 m 是不知道的。因此，使用式(4.8)就有困难。为此，构造函数 $u(x)$：当 $x=s$ 时 $u(x)=0$；当 $x\neq s$ 时

$$u(x)=\frac{f(x)}{f'(x)}$$

因 s 是 $f(x)$的 m 重零点，故 $u'(s)=\frac{1}{m}$，s 是 $u(x)$的单零点。求解方程 $u(x)=0$ 的 Newton 法迭代函数为

$$g(x)=x-\frac{u(x)}{u'(x)}=x-\frac{f(x)f'(x)}{[f'(x)]^2-f(x)f''(x)}$$

于是，迭代公式

$$x_{k+1}=x_k-\frac{f(x_k)f'(x_k)}{[f'(x_k)]^2-f(x_k)f''(x_k)}\quad(k=0,1,\cdots)\tag{4.9}$$

产生的序列$\{x_k\}$如果收敛于方程(4.1)的 m 重根 s，就至少是二阶收敛的。它的缺点是要计算 $f''(x)$，运算量稍大一些。

例如，已知方程

$$f(x) = x^4 - 1.4x^3 - 0.48x^2 + 1.408x - 0.512 = 0$$

有一个三重根 $s=0.8$，这里

$$f'(x) = 4x^3 - 4.2x^2 - 0.96x + 1.408$$

$$f''(x) = 12x^2 - 8.4x - 0.96$$

如果使用 Newton 法(4.7)进行迭代，并取初始值 $x_0=1$，则有

$$x_1 = 0.935\ 483\ 871, \quad x_2 = 0.891\ 352\ 317$$

$$x_3 = 0.861\ 384\ 032, \quad x_4 = 0.841\ 145\ 162$$

$$x_5 = 0.827\ 531\ 520, \quad x_6 = 0.818\ 400\ 189$$

$$x_7 = 0.812\ 287\ 422, \quad x_8 = 0.808\ 200\ 827$$

$$x_9 = 0.805\ 471\ 387, \quad x_{10} = 0.803\ 649\ 381$$

越往后收敛越慢。如果使用迭代公式(4.9)，也取初始值 $x_0=1$，则有

$$x_1 = 0.794\ 019\ 933$$

$$x_2 = 0.799\ 962\ 734$$

$$x_3 = 0.800\ 019\ 389$$

收敛得非常快。

4.1.7 割线法

Newton 法的一个明显缺点是对每一个 k 都要计算 $f'(x_k)$。导数的计算往往十分麻烦，尤其当$|f'(x_k)|$很小时，计算要很精确，否则会产生很大的舍入误差。本小节所阐述的割线法不须计算导数，虽然它的收敛速度低于 Newton 法，但高于一阶。因此，割线法在非线性方程的求解中得到广泛的应用。

在 Newton 法(4.7)中，用增量比$\dfrac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$替换导数$f'(x_k)$所成的迭代公式

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} \quad (k = 0,1,\cdots) \tag{4.10}$$

称为割线法(又称弦截法)，其中初始值 x_{-1}，x_0 预先给定。像简单迭代法那样，可用 x_{k-1}与 x_k 的接近程度控制迭代终止，也可用$|f(x_k)|<\eta$结束迭代，η是允许误差。

割线法(4.10)不仅用于求方程(4.1)的实数根，也能用于求方程(4.1)的复数根。割线法这个名称来自此方法求方程的实数根时的几何意义。设已获得方程(4.1)的两个近似根 x_{k-1} 和 x_k。过曲线 $y=f(x)$上的两点$(x_{k-1},f(x_{k-1}))$和$(x_k,f(x_k))$作曲线的割线，此割线的方程为

$$y - f(x_k) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}(x - x_k)$$

在上式中，令 $y=0$，并解出 x，所得的 x 就是迭代公式(4.10)中的 x_{k+1}。因此，x_{k+1}是割线与 x 轴相交所得交点的横坐标，见图 4-3。

现在讨论割线法(4.10)的收敛性。

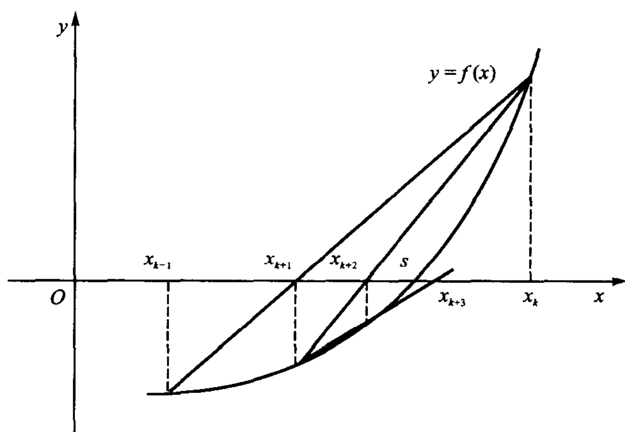


图 4-3　割线法

引理 4.1　设 $f(s)=0$，在 s 的某邻域 $[s-\delta,s+\delta]$ 内 $f''(x)$ 连续，$f'(x)\neq 0$，又设 x_{k-1}，$x_k\in[s-\delta,s+\delta]$ 且 x_{k-1}，x_k，s 互异，记 $e_k=s-x_k$，则有

$$e_{k+1}=s-x_{k+1}=e_k e_{k-1}\left[-\frac{f''(\eta_k)}{2f'(\xi_k)}\right] \tag{4.11}$$

其中 x_{k+1} 由公式(4.10)产生，η_k，ξ_k 在 $\min(x_{k-1},x_k,s)$ 与 $\max(x_{k-1},x_k,s)$ 之间。

证　由式(4.10)可得

$$e_{k+1}=s-x_{k+1}=\frac{f(x_k)e_{k-1}-f(x_{k-1})e_k}{f(x_k)-f(x_{k-1})}=$$

$$e_k e_{k-1}\frac{f(x_k)/e_k-f(x_{k-1})/e_{k-1}}{f(x_k)-f(x_{k-1})}=$$

$$-e_k e_{k-1}\frac{G(x_k)-G(x_{k-1})}{f(x_k)-f(x_{k-1})}$$

其中

$$G(x)=\begin{cases}\dfrac{f(x)-f(s)}{x-s}, & x\neq s\\ f'(s), & x=s\end{cases}$$

由柯西中值定理，得

$$e_{k+1}=-e_k e_{k-1}\frac{G'(\xi_k)}{f'(\xi_k)}$$

其中 ξ_k 在 x_{k-1} 与 x_k 之间。若 $\xi_k\neq s$，则

$$G'(\xi_k)=\frac{f'(\xi_k)(\xi_k-s)-f(\xi_k)+f(s)}{(\xi_k-s)^2}=\frac{f''(\eta_k)}{2}$$

其中 η_k 在 ξ_k 与 s 之间。若 $\xi_k=s$，则

$$G'(\xi_k)=G'(s)=\frac{f''(s)}{2}$$

于是有

$$e_{k+1}=e_k e_{k-1}\left[-\frac{f''(\eta_k)}{2f'(\xi_k)}\right]$$

其中 η_k 和 ξ_k 在 $\min(x_{k-1},x_k,s)$ 与 $\max(x_{k-1},x_k,s)$ 之间。

证毕。

在下面的叙述中，记 $I_\alpha=[s-\alpha,s+\alpha]$。

定理 4.8 设 $f(s)=0$，在 s 的某邻域内 $f''(x)$ 连续且 $f'(x)\neq 0$，则存在 $\varepsilon>0$，当 $x_{-1},x_0\in I_\varepsilon$ 时，由割线法(4.10)产生的序列 $\{x_k\}$ 收敛于 s，且收敛速度的阶至少为 1.618。

证 由于 $f'(x)$ 和 $f''(x)$ 在 s 的某邻域内连续且 $f'(x)\neq 0$，所以必存在 $\beta>0$ 及相应的数 $M_\beta>0$，使得

$$\left|\frac{f''(x)}{f'(\tilde{x})}\right|\leqslant M_\beta,\quad x,\tilde{x}\in I_\beta$$

选择 $\varepsilon>0$，使其满足 $\varepsilon M_\beta\equiv K<1$ 且 $\varepsilon<\beta$，因而 $I_\varepsilon\subset I_\beta$。令 $x_{-1},x_0\in I_\varepsilon$，那么，根据引理 4.1，得不等式

$$|e_1|<|e_{-1}e_0|M_\beta\leqslant\varepsilon^2M_\beta=\varepsilon K<\varepsilon$$

$$|e_2|<|e_0e_1|M_\beta<\varepsilon^2KM_\beta=\varepsilon K^2<\varepsilon$$

今假定 $|e_{k-1}|<\varepsilon K^{k-1}$ 和 $|e_k|<\varepsilon K^k$ 成立，则有

$$|e_{k+1}|<|e_{k-1}e_k|M_\beta<\varepsilon^2K^kM_\beta=\varepsilon K^{k+1}<\varepsilon$$

由归纳法原理可知，对任意的 k，均有

$$|e_k|<\varepsilon K^k,\quad x_k\in I_\varepsilon$$

因此，$\lim\limits_{k\to\infty}e_k=0$，即 $\{x_k\}$ 收敛于 s。

设

$$\lim_{k\to\infty}\frac{|e_{k+1}|}{|e_k|^r}=c$$

r 和 c 是待定的正实数。由上式得

$$|e_{k+1}|=(c+\alpha_k)|e_k|^r$$

其中 $\alpha_k\to 0(k\to\infty)$。由此又得

$$|e_{k-1}|=|e_k|^{\frac{1}{r}}(c+\alpha_{k-1})^{-\frac{1}{r}}$$

根据引理 4.1 的式(4.11)，可得

$$(c+\alpha_k)|e_k|^r=|e_k||e_k|^{\frac{1}{r}}(c+\alpha_{k-1})^{-\frac{1}{r}}\left|\frac{f''(\eta_k)}{2f'(\xi_k)}\right|$$

$$|e_k|^{r-(1+\frac{1}{r})}=(c+\alpha_k)^{-1}(c+\alpha_{k-1})^{-\frac{1}{r}}\left|\frac{f''(\eta_k)}{2f'(\xi_k)}\right| \tag{4.12}$$

令 $k\to\infty$，等式(4.12)右边的极限存在且为

$$c^{-(1+\frac{1}{r})}\left|\frac{f''(s)}{2f'(s)}\right|=t$$

若 $f''(s)\neq 0$，则 $t\neq 0$，那么，要使等式(4.12)左边的极限也不为零，只能

$$r-\left(1+\frac{1}{r}\right)=0$$

即

$$r^2-r-1=0$$

因而 $t=1$。这时，正数 r 和正数 c 分别为

$$r=\frac{1}{2}(1+\sqrt{5})=1.618\cdots$$

$$c=\left|\frac{f''(s)}{2f'(s)}\right|^{\frac{1}{r}}=\left|\frac{f''(s)}{2f'(s)}\right|^{0.618\cdots}$$

若 $f''(s)=0$，则 $t=0$，要使等式(4.12)左边的极限也为零，必须

$$r-\left(1+\frac{1}{r}\right)>0$$

这时，正数 $r>\frac{1}{2}(1+\sqrt{5})=1.618\cdots$。

综上所述，可知序列 $\{x_k\}$ 的收敛速度的阶至少是 1.618。

证毕。

例 4 用割线法求方程 $x-\ln x=2$ 在区间 $(2,\infty)$ 内的根，要求 $\frac{|x_k-x_{k-1}|}{|x_k|}<10^{-8}$。

解 前已分析，所求根位于区间(2,4)内。迭代公式为

$$x_{k+1}=x_k-\frac{(x_k-x_{k-1})}{f(x_k)-f(x_{k-1})}f(x_k)\quad(k=0,1,\cdots)$$

其中 $f(x)=x-\ln x-2$。取 $x_{-1}=2$，$x_0=4$，迭代结果见表 4-5。x_5 已达到精度要求，故方程的根 $s\approx3.146\ 193\ 221$。

表 4-5 例 4 计算结果

k	x_k	$f(x_k)$
−1	2.000 000 000	−0.693 147 180
0	4.000 000 000	0.613 705 638
1	3.060 788 438	−0.057 884 104
2	3.141 738 781	−0.003 037 617
3	3.146 222 134	0.000 019 723
4	3.146 193 211	-0.6×10^{-8}
5	3.146 193 221	0.0

比较例 1、例 3 和例 4 也可看出，求解同一个方程的同一个根，为达到相同精度，例 1 所用的简单迭代法由于只有线性收敛速度，因而迭代次数最多；例 3 用的是 Newton 法，由于有二阶收敛速度，因而迭代次数最少；例 4 用的是割线法，迭代次数比例 1 少但比例 3 多，这是由于割线法的收敛速度的阶介于前两者之间。

4.1.8 单点割线法

在割线法(4.10)中，用固定点 $(x_0,f(x_0))$ 代替 $(x_{k-1},f(x_{k-1}))$，也就是点 $(x_0,f(x_0))$ 永远是割线上的一点，就得到新的迭代公式

$$x_{k+1}=x_k-\frac{x_k-x_0}{f(x_k)-f(x_0)}f(x_k)\quad(k=1,2,\cdots)\tag{4.13}$$

称迭代公式(4.13)为单点割线法(或单点弦截法)，它属于简单迭代法。

单点割线法有类似于定理 4.7 的收敛性定理。

定理 4.9 设函数 $f(x)$ 在区间 $[a,b]$ 上存在二阶连续导数，且满足条件：

(1) $f(a)f(b)<0$；

(2) $f''(x)$在区间$[a,b]$上不变号；

(3) 当 $x\in[a,b]$时，$f'(x)\neq 0$；

(4) $x_0,x_1\in[a,b]$且 $f(x_0)f''(x_0)>0, f(x_0)f(x_1)<0$。

则由单点割线法(4.13)产生的序列$\{x_k\}$单调收敛于方程(4.1)在$[a,b]$内唯一的根 s，并且收敛速度是一阶的。

证 条件(1)，(2)共有四种情况，现在仅就其中一种情况进行证明，其余情况的证明类似。

设 $f(a)<0, f(b)>0$，当 $x\in[a,b]$时，$f''(x)\leqslant 0$。仿照定理 4.7 的证明方法可证得方程(4.1)在区间$[a,b]$内有唯一的根 s，并且 $f(x)$在$[a,b]$上严格单调增大(见图 4-4)。

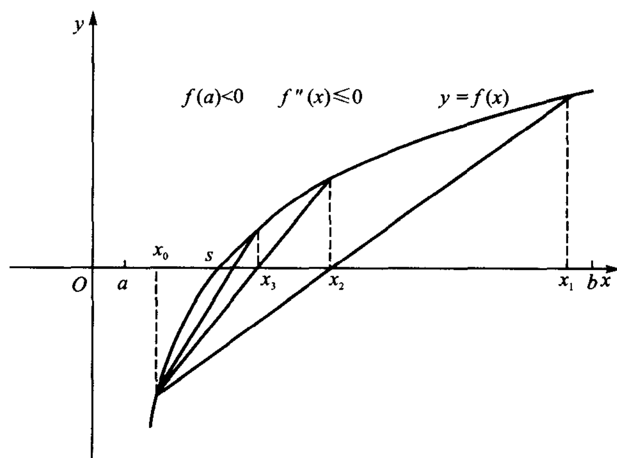


图 4-4 定理 4.9 的一种情形

由条件(4)可知，$f(x_0)<0, f(x_1)>0$，因而 $x_0<s<x_1$。由式(4.13)得

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1) < x_1$$

下面证明 $s\leqslant x_2$，因为 $f(x)$单调增大，故只须证明 $f(x_2)\geqslant 0$。为此，构造函数

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0}, \quad x \in (x_0, b)$$

由于

$$g'(x) = \frac{f'(x)(x - x_0) - f(x) + f(x_0)}{(x - x_0)^2} = \frac{f''(\xi)}{2} \leqslant 0, \quad \xi \in (x_0, x)$$

所以，$g(x)$在区间(x_0,b)上单调减小。因 $x_2<x_1$，故有 $g(x_2)\geqslant g(x_1)$，即

$$\frac{f(x_2) - f(x_0)}{x_2 - x_0} \geqslant \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$f(x_2) \geqslant f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_2 - x_0) = 0$$

由此证明了 $s\leqslant x_2$。于是有 $x_0<s\leqslant x_2<x_1$。

一般地，由 $x_0<s\leqslant x_k$ 同样可证明由式(4.13)得到的 x_{k+1}满足

$$x_0 < s \leqslant x_{k+1} \leqslant x_k$$

因此由式(4.13)产生的序列$\{x_k\}$单调减小有下界，因而有极限 $\tilde{s}$。对式(4.13)两边取极限就

得到 $f(\tilde{s})=0$。但方程 $f(x)=0$ 在$[a,b]$内的根 s 是唯一的，所以 $\tilde{s}=s$。

迭代公式(4.13)的迭代函数为

$$\varphi(x)=x-\frac{x-x_0}{f(x)-f(x_0)}f(x)=\frac{x_0 f(x)-xf(x_0)}{f(x)-f(x_0)}$$

由

$$\varphi'(x)=\frac{f(x_0)[f(x_0)-f(x)+(x-x_0)f'(x)]}{[f(x)-f(x_0)]^2}$$

得

$$\varphi'(s)=\frac{1}{f(x_0)}[f(x_0)+(s-x_0)f'(s)]=\frac{f''(\xi)}{2f(x_0)}(x_0-s)^2,\quad \xi\in(x_0,s)$$

一般 $f''(\xi)\neq 0$，故迭代公式(4.13)产生的序列$\{x_k\}$是一阶收敛的。

证毕。

例 5　用单点割线法求方程 $x-\ln x=2$ 在区间$(2,\infty)$内的根，要求$\frac{|x_k-x_{k-1}|}{|x_k|}<10^{-8}$。

解　$f(x)=x-\ln x-2$ 满足

(1) $f(2)f(4)<0$；

(2) 当 $x\in[2,4]$时，$f''(x)>0$；

(3) 在区间$[2,4]$上 $f'(x)\neq 0$；

(4) $f(4)f''(4)>0$。

所以，选 $x_0=4, x_1=2$，由单点割线法(4.13)产生的序列$\{x_k\}$必收敛于方程在$[2,4]$内的根 s。迭代结果见表 4-6。x_8 已满足精度要求，故方程的根为 $s\approx 3.146\ 193\ 219$。

表 4-6　例 5 计算结果

k	x_k	$f(x_k)$
0	4.000 000 000	0.613 705 638
1	2.000 000 000	−0.693 147 180
2	3.060 788 439	−0.057 884 103
3	3.141 738 781	−0.003 037 617
4	3.145 965 936	−0.000 155 040
5	3.146 181 637	$-0.790\ 2\times 10^{-5}$
6	3.146 192 630	-0.402×10^{-6}
7	3.146 193 191	-0.20×10^{-7}
8	3.146 193 219	-0.1×10^{-8}

4.2　非线性方程组的迭代解法

4.2.1　一般概念

含 n 个方程的 n 元非线性方程组的一般形式是

$$\begin{cases} f_1(x_1,x_2,\cdots,x_n)=0 \\ f_2(x_1,x_2,\cdots,x_n)=0 \\ \quad\vdots \\ f_n(x_1,x_2,\cdots,x_n)=0 \end{cases} \tag{4.14}$$

其中 $f_i(i=1,2,\cdots,n)$是定义在区域 $D\subset\mathbf{R}^n$ 上的 n 元实值函数，且 f_i 中至少有一个是非线性函数。令

$$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$$
$$\boldsymbol{F}(\boldsymbol{x})=(f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\cdots,f_n(\boldsymbol{x}))^{\mathrm{T}}$$

则方程组(4.14)可表示为向量形式

$$\boldsymbol{F}(\boldsymbol{x})=\boldsymbol{0} \tag{4.15}$$

其中 $\boldsymbol{F}:D\subset\mathbf{R}^n\to\mathbf{R}^n$，即 $\boldsymbol{F}$ 是定义在区域 $D\subset\mathbf{R}^n$ 上且是 n 维实向量值函数。若存在 $\boldsymbol{x}^*\in D$，使 $\boldsymbol{F}(\boldsymbol{x}^*)=\boldsymbol{0}$，则称 $\boldsymbol{x}^*$ 是方程组(4.15)的解。

研究非线性方程组(4.15)的解的存在性和有效解法已有很多成果，本节只能简要介绍其中几种迭代解法。为此先介绍有关概念。

定义 设 $f:D\subset\mathbf{R}^n\to\mathbf{R}^1$，$\boldsymbol{x}\in\mathrm{int}(D)$（即 $\boldsymbol{x}$ 是 D 的内点），若存在向量 $\boldsymbol{l}(\boldsymbol{x})\in\mathbf{R}^n$，使极限

$$\lim_{\boldsymbol{h}\to 0}\frac{f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{l}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}}{\|\boldsymbol{h}\|}=0 \tag{4.16}$$

成立，则称 f 在 $\boldsymbol{x}$ 处可微，向量 $\boldsymbol{l}(\boldsymbol{x})$称为 f 在 $\boldsymbol{x}$ 处的导数，记为$f'(\boldsymbol{x})=\boldsymbol{l}(\boldsymbol{x})$；若 D 是开区域且 f 在 D 内每点处都可微，则称 f 在 D 可微。

定理 4.10 若 $f:D\subset\mathbf{R}^n\to\mathbf{R}^1$ 在 $\boldsymbol{x}\in\mathrm{int}(D)$处可微，则 f 在 $\boldsymbol{x}$ 处关于各自变量的偏导数 $\dfrac{\partial f(\boldsymbol{x})}{\partial x_j}(j=1,2,\cdots,n)$存在，且有

$$f'(\boldsymbol{x})=\left(\frac{\partial f(\boldsymbol{x})}{\partial x_1},\frac{\partial f(\boldsymbol{x})}{\partial x_2},\cdots,\frac{\partial f(\boldsymbol{x})}{\partial x_n}\right)^{\mathrm{T}}$$

证 记 $\boldsymbol{l}(\boldsymbol{x})=(l_1(\boldsymbol{x}),l_2(\boldsymbol{x}),\cdots,l_n(\boldsymbol{x}))^{\mathrm{T}}$，取 $\boldsymbol{h}=h\boldsymbol{e}_j$（实数 $h\neq 0$，$\boldsymbol{e}_j$ 是 n 维基本单位向量），由于(4.16)成立，故有

$$\lim_{h\to 0}\frac{f(\boldsymbol{x}+h\boldsymbol{e}_j)-f(\boldsymbol{x})-l_j(\boldsymbol{x})h}{h}=0 \quad (j=1,2,\cdots,n)$$

因而

$$l_j(\boldsymbol{x})=\lim_{h\to 0}\frac{f(\boldsymbol{x}+h\boldsymbol{e}_j)-f(\boldsymbol{x})}{h}=\frac{\partial f(\boldsymbol{x})}{\partial x_j} \quad (j=1,2,\cdots,n)$$

存在，且有

$$f'(\boldsymbol{x})=\boldsymbol{l}(\boldsymbol{x})=\left(\frac{\partial f(\boldsymbol{x})}{\partial x_1},\frac{\partial f(\boldsymbol{x})}{\partial x_2},\cdots,\frac{\partial f(\boldsymbol{x})}{\partial x_n}\right)^{\mathrm{T}}$$

证毕。

f 在 $\boldsymbol{x}$ 处的导数 $f'(\boldsymbol{x})$又称为 f 在 $\boldsymbol{x}$ 处的梯度，又可记为 $\mathbf{grad}\ f(\boldsymbol{x})$和$\nabla f(\boldsymbol{x})$。

定义 设 $\boldsymbol{F}:D\subset\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{x}\in\mathrm{int}(D)$，若存在矩阵 $\boldsymbol{A}(\boldsymbol{x})\in\mathbf{R}^{n\times n}$，使极限

$$\lim_{\boldsymbol{h}\to 0}\frac{\|\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h})-\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{A}(\boldsymbol{x})\boldsymbol{h}\|}{\|\boldsymbol{h}\|}=0 \tag{4.17}$$

成立，则称 $\boldsymbol{F}$ 在 $\boldsymbol{x}$ 处可微，矩阵 $\boldsymbol{A}(\boldsymbol{x})$称为 $\boldsymbol{F}$ 在 $\boldsymbol{x}$ 处的导数，记为 $\boldsymbol{F}'(\boldsymbol{x})=\boldsymbol{A}(\boldsymbol{x})$；若 D 是开区域且 $\boldsymbol{F}$ 在 D 内每点处都可微，则称 $\boldsymbol{F}$ 在 D 可微。

定理 4.11 设 $\boldsymbol{F}:D\subset\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{F}$ 在 $\boldsymbol{x}\in\mathrm{int}(D)$处可微的充分必要条件是 $\boldsymbol{F}$ 的所有分量 f_i $(i=1,2,\cdots,n)$在 $\boldsymbol{x}$ 处可微；若 $\boldsymbol{F}$ 在 $\boldsymbol{x}$ 处可微，则

$$F'(x)=\left[\frac{\partial f_i(x)}{\partial x_j}\right]_{n\times n}=\begin{bmatrix}\frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_n}\\ \vdots & \vdots & & \vdots\\ \frac{\partial f_n(x)}{\partial x_1} & \frac{\partial f_n(x)}{\partial x_2} & \cdots & \frac{\partial f_n(x)}{\partial x_n}\end{bmatrix}$$

证　由于 $F(x)=(f_1(x),\cdots,f_n(x))^{\mathrm{T}}$，所以，存在向量 $l_i(x)\in\mathbf{R}^n$ 使极限

$$\lim_{h\to 0}\frac{f_i(x+h)-f_i(x)-l_i(x)^{\mathrm{T}}h}{\|h\|}=0\quad(i=1,2,\cdots,n)$$

成立与存在矩阵 $A(x)\in\mathbf{R}^{n\times n}$ 使极限(4.17)成立是等价的，并且 $A(x)=(l_1(x),\cdots,l_n(x))^{\mathrm{T}}$，即 $f_i(i=1,2,\cdots,n)$ 在 x 处可微是 F 在 x 处可微的充分必要条件。又根据定理 4.10，当 F 在 x 处可微时，

$$F'(x)=A(x)=\left[\frac{\partial f_i(x)}{\partial x_j}\right]_{n\times n}$$

证毕。

矩阵 $\left[\frac{\partial f_i(x)}{\partial x_j}\right]_{n\times n}$ 称为 F 在 x 处的 Jacobi 矩阵。

定理 4.12　设 $F:D\subset\mathbf{R}^n\to\mathbf{R}^n$，

(1) 若 F 在 $x\in\mathrm{int}(D)$ 处的 Jacobi 矩阵存在且连续，则 F 在 x 处可微，并称 F 在 x 处连续可微，且 $F'(x)=\left[\frac{\partial f_i(x)}{\partial x_j}\right]_{n\times n}$。

(2) 若 F 在 $x\in\mathrm{int}(D)$ 处可微，则 F 在 x 处连续。

(3) 若 F 在开区域 D 内可微，$D_0\subset D$ 为开凸域，则对任意的 $x\in D_0$ 和 $x+h\in D_0$，等式

$$F(x+h)-F(x)=\begin{bmatrix}f'_1(x+\theta_1 h)^{\mathrm{T}}\\ f'_2(x+\theta_2 h)^{\mathrm{T}}\\ \vdots\\ f'_n(x+\theta_n h)^{\mathrm{T}}\end{bmatrix}h \tag{4.18}$$

成立，其中 $0<\theta_i<1\ (i=1,2,\cdots,n)$。

证明从略。

定义　若 $F:D\subset\mathbf{R}^n\to\mathbf{R}^n$ 的各个分量 $f_i(x)(i=1,2,\cdots,n)$ 的二阶偏导数在 $x\in\mathrm{int}(D)$ 处连续，则称 $F(x)$ 在 x 处二次连续可微。

定义　设向量序列 $\{x_k\}$ 收敛于 x^*，$e_k=x^*-x_k\neq 0(k=0,1,\cdots)$，如果存在常数 $r\geqslant 1$ 和常数 $c>0$，使得极限

$$\lim_{k\to\infty}\frac{\|e_{k+1}\|}{\|e_k\|^r}=c$$

成立，或者使得当 $k\geqslant K$(某个正整数)时，

$$\|e_{k+1}\|\leqslant c\|e_k\|^r$$

成立，则称序列 $\{x_k\}$ 收敛于 x^* 具有 r 阶收敛速度，简称 $\{x_k\}$ 是 r 阶收敛的，c 称为收敛因子。

当 $r=1$ 时，称序列 $\{x_k\}$ 是线性收敛的，此时必有 $0<c\leqslant 1$；当 $r>1$ 时，称序列 $\{x_k\}$ 是超线性收敛的；当 $r=2$ 时，称序列 $\{x_k\}$ 是平方收敛的。

4.2.2 简单迭代法

把方程组(4.15)改写成与之等价的形式

$$x = G(x) \tag{4.19}$$

其中 $G: D \subset \mathbf{R}^n \to \mathbf{R}^n$。若 $x^* \in D$ 满足 $x^* = G(x^*)$，则称 x^* 为函数$G(x)$的不动点。因此，$G(x)$的不动点就是方程组(4.15)的解，求方程组(4.15)的解就转化为求函数 $G(x)$的不动点。

适当选取初始向量 $x^{(0)} \in D$，利用方程组(4.19)的形式，构成迭代公式

$$x^{(k+1)} = G(x^{(k)}) \quad (k = 0,1,\cdots) \tag{4.20}$$

迭代公式(4.20)称为求解方程组(4.19)的简单迭代法，又称为不动点迭代法，$G(x)$称为迭代函数。

定理 4.13(压缩映像原理) 设 $G: D \subset \mathbf{R}^n \to \mathbf{R}^n$ 在闭区域 $D_0 \subset D$ 上满足两个条件：

(1) G 把 D_0 映入它自身，即 $G(D_0) \subset D_0$；

(2) G 在 D_0 上是压缩映射，即存在常数 $L \in (0,1)$，使对任意的 $x, y \in D_0$，有

$$\| G(x) - G(y) \| \leqslant L \| x - y \| \tag{4.21}$$

则有下列结论：

(1) 对任取的 $x^{(0)} \in D_0$，由迭代公式(4.20)产生的序列$\{x^{(k)}\} \subset D_0$，且收敛于方程组(4.19)在 D_0 内的唯一解 x^*；

(2) 成立误差估计式

$$\| x^* - x^{(k)} \| \leqslant \frac{L^k}{1-L} \| x^{(1)} - x^{(0)} \| \tag{4.22}$$

$$\| x^* - x^{(k)} \| \leqslant \frac{L}{1-L} \| x^{(k)} - x^{(k-1)} \| \tag{4.23}$$

证 (1) 由于 $x^{(0)} \in D_0$ 以及条件(1)，可知迭代公式(4.20)产生的序列有意义且$\{x^{(k)}\} \subset D_0$。又由条件(2)得

$$\| x^{(k+1)} - x^{(k)} \| = \| G(x^{(k)}) - G(x^{(k-1)}) \| \leqslant L \| x^{(k)} - x^{(k-1)} \|$$

当 $m \geqslant 1$ 时，有

$$\begin{aligned} \| x^{(k+m)} - x^{(k)} \| &\leqslant \sum_{i=1}^{m} \| x^{(k+i)} - x^{(k+i-1)} \| \leqslant \\ & \sum_{i=1}^{m} L^{k+i-1} \| x^{(1)} - x^{(0)} \| \leqslant \\ & \frac{L^k}{1-L} \| x^{(1)} - x^{(0)} \| \end{aligned} \tag{4.24}$$

因 $0 < L < 1$，根据 Cauchy(柯西)收敛原理可知序列$\{x^{(k)}\}$收敛。又因 D_0 是闭区域，故存在 $x^* \in D_0$，使$\lim\limits_{k \to \infty} x^{(k)} = x^*$。由条件(2)又可知 $G(x)$在 D_0 上连续，因而有

$$x^* = \lim_{k \to \infty} x^{(k+1)} = \lim_{k \to \infty} G(x^{(k)}) = G(x^*)$$

即 x^* 是方程组(4.19)的解。

设 $x^*, y^* \in D_0$ 是方程组(4.19)的两个不同的解，则有

$$\begin{aligned} \| x^* - y^* \| &= \| G(x^*) - G(y^*) \| \leqslant \\ & L \| x^* - y^* \| < \| x^* - y^* \| \end{aligned}$$

所出现的矛盾证实 $\boldsymbol{x}^*$ 是方程组(4.19)在 D_0 内的唯一解。

(2) 由式(4.24),令 $m\to\infty$,得到式(4.22)。又当 $m\geqslant 1$ 时,有

$$\begin{aligned}\|\boldsymbol{x}^{(k+m)}-\boldsymbol{x}^{(k)}\| \leqslant &\sum_{i=1}^{m}\|\boldsymbol{x}^{(k+i)}-\boldsymbol{x}^{(k+i-1)}\| \leqslant \\ &\sum_{i=1}^{m}L^i\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\| \leqslant \\ &\frac{L}{1-L}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|\end{aligned}$$

再令 $m\to\infty$,就得到式(4.23)。

证毕。

实际计算时,可预先给定精度水平 $\varepsilon>0$,当迭代序列满足 $\dfrac{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|}{\|\boldsymbol{x}^{(k)}\|}\leqslant\varepsilon$ 时停止迭代,取当前的 $\boldsymbol{x}^{(k)}$ 作为方程组(4.19)的近似解。也可利用式(4.22),根据给定的误差限 $\eta>0$,预先确定迭代次数 k,使近似解 $\boldsymbol{x}^{(k)}$ 满足 $\|\boldsymbol{x}^*-\boldsymbol{x}^{(k)}\|\leqslant\eta$。

在定理 4.13 的条件下,简单迭代法(4.20)产生的序列 $\{\boldsymbol{x}^{(k)}\}$ 满足

$$\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\| = \|\boldsymbol{G}(\boldsymbol{x}^{(k)})-\boldsymbol{G}(\boldsymbol{x}^*)\| \leqslant L\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|$$

其中 $0<L<1$,因此,序列 $\{\boldsymbol{x}^{(k)}\}$ 是线性收敛的。

若 $\boldsymbol{G}(\boldsymbol{x})$ 是向量 $\boldsymbol{x}\in\mathbf{R}^n$ 的线性函数,即 $\boldsymbol{G}(\boldsymbol{x})=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{d}$,其中 $\boldsymbol{B}\in\mathbf{R}^{n\times n}$ 是常数矩阵,$\boldsymbol{d}\in\mathbf{R}^n$ 是常向量,则迭代公式(4.20)成为

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{d}\quad(k=0,1,\cdots)\tag{4.25}$$

它是求解线性方程组 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{d}$ 的迭代法。此时,$\boldsymbol{G}(\mathbf{R}^n)\subset\mathbf{R}^n$,且对任意的 $\boldsymbol{x},\boldsymbol{y}\in\mathbf{R}^n$,有

$$\|\boldsymbol{G}(\boldsymbol{x})-\boldsymbol{G}(\boldsymbol{y})\| = \|\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{y})\| \leqslant \|\boldsymbol{B}\|\,\|\boldsymbol{x}-\boldsymbol{y}\|$$

因此,根据定理 4.13 可知,当 $\|\boldsymbol{B}\|<1$ 时,对任取的 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$,迭代公式(4.25)产生的序列 $\{\boldsymbol{x}^{(k)}\}$ 必收敛于方程组 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{d}$ 的唯一解 $\boldsymbol{x}^*$。这个结果与定理 2.9 完全一致。

下面给出简单迭代法(4.20)的局部收敛性定理。

定理 4.14　设 $\boldsymbol{G}:D\subset\mathbf{R}^n\to\mathbf{R}^n$,$\boldsymbol{x}^*\in\mathrm{int}(D)$ 是方程组(4.19)的解,$\boldsymbol{G}$ 在 $\boldsymbol{x}^*$ 处可微。若 $\boldsymbol{G}'(\boldsymbol{x}^*)$ 的谱半径 $\rho(\boldsymbol{G}'(\boldsymbol{x}^*))<1$,则存在开球 $D_0=\{\boldsymbol{x}\mid\|\boldsymbol{x}-\boldsymbol{x}^*\|<\delta,\delta>0\}\subset D$,使对任意的 $\boldsymbol{x}^{(0)}\in D_0$,由迭代法(4.20)产生的序列 $\{\boldsymbol{x}^{(k)}\}\subset D_0$ 且收敛于 $\boldsymbol{x}^*$。

证明从略。

对于线性方程组 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{d}$,此时 $\boldsymbol{G}(\boldsymbol{x})=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{d}$ 在 $\mathbf{R}^n$ 中处处可微,且对任意的 $\boldsymbol{x}\in\mathbf{R}^n$,$\boldsymbol{G}'(\boldsymbol{x})=\boldsymbol{B}$是常数矩阵。根据定理 4.14,若 $\boldsymbol{B}$ 的谱半径 $\rho(\boldsymbol{B})<1$,则对任取的 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$,由迭代法(4.25)产生的序列 $\{\boldsymbol{x}^{(k)}\}$ 必收敛于方程组 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{d}$ 的解 $\boldsymbol{x}^*$。可见,对于线性方程组来说,定理 4.14 成为全局收敛性定理。此外,从定理 2.8 知道,$\rho(\boldsymbol{B})<1$ 不仅是迭代法(4.25)收敛的充分条件,而且也是必要条件。但是,对于非线性方程组(4.19),条件 $\rho(\boldsymbol{G}'(\boldsymbol{x}^*))<1$ 只是迭代法(4.20)收敛于 $\boldsymbol{x}^*$ 的充分条件,而不是必要条件,其原因是 $\boldsymbol{G}'(\boldsymbol{x})$不是常数矩阵而是依赖于变向量 $\boldsymbol{x}$ 的函数矩阵,这正是线性与非线性的本质区别。

例 6　用简单迭代法求解下列方程组

$$\begin{cases}3x_1-\cos x_1-\sin x_2=0\\4x_2-\sin x_1-\cos x_2=0\end{cases}$$

要求满足精度$\frac{\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \|_\infty}{\| \boldsymbol{x}^{(k)} \|_\infty} \leqslant 10^{-12}$。

解 把方程组写成等价形式：

$$\begin{cases} x_1 = \frac{1}{3}(\cos x_1 + \sin x_2) \\ x_2 = \frac{1}{4}(\sin x_1 + \cos x_2) \end{cases}$$

记

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \boldsymbol{G}(\boldsymbol{x}) = \begin{bmatrix} \frac{1}{3}(\cos x_1 + \sin x_2) \\ \frac{1}{4}(\sin x_1 + \cos x_2) \end{bmatrix}$$

则有 $\boldsymbol{G}(\mathbf{R}^2) \subset \mathbf{R}^2$，且对任意的 $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^2$ 有

$$\boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{G}(\boldsymbol{y}) = \begin{bmatrix} \frac{1}{3}(\cos x_1 - \cos y_1 + \sin x_2 - \sin y_2) \\ \frac{1}{4}(\sin x_1 - \sin y_1 + \cos x_2 - \cos y_2) \end{bmatrix} = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})$$

其中 $\boldsymbol{A} = \begin{bmatrix} -\frac{1}{3}\sin \xi_1 & \frac{1}{3}\cos \xi_2 \\ \frac{1}{4}\cos \eta_1 & -\frac{1}{4}\sin \eta_2 \end{bmatrix}$；$\xi_1, \eta_1$ 在 x_1 与 y_1 之间；ξ_2, η_2 在 x_2 与 y_2 之间。于是有

$$\| \boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{G}(\boldsymbol{y}) \|_2 \leqslant \| \boldsymbol{A} \|_F \| \boldsymbol{x} - \boldsymbol{y} \|_2 \leqslant \sqrt{\frac{25}{72}} \| \boldsymbol{x} - \boldsymbol{y} \|_2$$

因此，对任取的 $\boldsymbol{x}^{(0)} \in \mathbf{R}^2$，迭代公式

$$\begin{cases} x_1^{(k+1)} = \frac{1}{3}(\cos x_1^{(k)} + \sin x_2^{(k)}) \\ x_2^{(k+1)} = \frac{1}{4}(\sin x_1^{(k)} + \cos x_2^{(k)}) \\ \qquad (k = 0, 1, \cdots) \end{cases}$$

产生的序列 $\{\boldsymbol{x}^{(k)}\}$ 必收敛于所给方程组在 $\mathbf{R}^2$ 中的唯一解 $\boldsymbol{x}^*$。取 $x_1^{(0)} = x_2^{(0)} = 1$，则当 $k = 28$ 时满足精度要求，得

$$x_1 \approx 0.415\ 169\ 427\ 139$$
$$x_2 \approx 0.336\ 791\ 217\ 025$$

4.2.3 Newton 法

设方程组(4.15)存在解 $\boldsymbol{x}^* \in \mathrm{int}(D)$，在 $\boldsymbol{x}^*$ 的某个开邻域 $D_0 = \{\boldsymbol{x} \mid \| \boldsymbol{x} - \boldsymbol{x}^* \| < \delta, \delta > 0\} \subset D$ 内 $\boldsymbol{F}(\boldsymbol{x})$ 可微。又设 $\boldsymbol{x}^{(k)} \in D_0$ 是方程组(4.15)的第 k 个近似解。由一阶 Taylor 公式，可得

$$f_i(\boldsymbol{x}) \approx f_i(\boldsymbol{x}^{(k)}) + \sum_{j=1}^{n} \frac{\partial f_i(\boldsymbol{x}^{(k)})}{\partial x_j}(x_j - x_j^{(k)}) \quad (i = 1, 2, \cdots, n)$$

今用线性方程组

$$f_i(\boldsymbol{x}^{(k)}) + \sum_{j=1}^{n} \frac{\partial f_i(\boldsymbol{x}^{(k)})}{\partial x_j}(x_j - x_j^{(k)}) = 0 \quad (i = 1, 2, \cdots, n)$$

即
$$F'(x^{(k)})(x-x^{(k)})=-F(x^{(k)}) \tag{4.26}$$
近似代替非线性方程组(4.15)，用线性方程组(4.26)的解作为非线性方程组(4.15)第 $k+1$ 个近似解，由此得到迭代公式
$$x^{(k+1)}=x^{(k)}-F'(x^{(k)})^{-1}F(x^{(k)}) \quad (k=0,1,\cdots) \tag{4.27}$$
称迭代公式(4.27)为求解非线性方程组(4.15)的 Newton 法。特殊地，当 $n=1$ 时，迭代公式(4.27)就成为求解一元非线性方程 $f_1(x_1)=0$ 的 Newton 法
$$x_1^{(k+1)}=x_1^{(k)}-\frac{f_1(x_1^{(k)})}{f'_1(x_1^{(k)})} \quad (k=0,1,\cdots)$$

定理 4.15　设 $x^*\in \mathrm{int}(D)$ 是方程组(4.15)的解，$F:D\subset \mathbf{R}^n\to\mathbf{R}^n$ 在包含 x^* 的某个开区域 $S\subset D$ 内连续可微，且 $F'(x^*)$ 非奇异，则存在闭球 $D_0=\{x \mid \|x-x^*\|\leqslant\delta,\delta>0\}\subset S$，使对任意的 $x_0\in D_0$，由 Newton 法(4.27)产生的序列 $\{x^{(k)}\}\subset D_0$ 且超线性收敛于 x^*；若更有 $F(x)$ 在域 S 内二次连续可微，则序列 $\{x^{(k)}\}$ 至少是平方收敛的。

证明从略。

Newton 法求方程组(4.15)的解 x^*，可采用如下的算法：

(1) 在 x^* 附近选取 $x^{(0)}\in D$，给定精度水平 $\varepsilon>0$ 和最大迭代次数 M。

(2) 对于 $k=0,1,\cdots,M$ 执行

① 计算 $F(x^{(k)})$ 和 $F'(x^{(k)})$。

② 求解关于 $\Delta x^{(k)}$ 的线性方程组
$$F'(x^{(k)})\Delta x^{(k)}=-F(x^{(k)})$$

③ 若 $\|\Delta x^{(k)}\|/\|x^{(k)}\|\leqslant\varepsilon$，则取 $x^*\approx x^{(k)}$，并停止计算；否则转④。

④ 计算 $x^{(k+1)}=x^{(k)}+\Delta x^{(k)}$。

⑤ 若 $k<M$，则继续；否则，输出 M 次迭代不成功的信息，并停止计算。

例 7　试用 Newton 法求解例 6 的方程组，精度要求与例 6 相同。

解　Newton 法迭代公式为
$$\begin{cases}\begin{bmatrix}3+\sin x_1^{(k)} & -\cos x_2^{(k)}\\ -\cos x_1^{(k)} & 4+\sin x_2^{(k)}\end{bmatrix}\begin{bmatrix}\Delta x_1^{(k)}\\ \Delta x_2^{(k)}\end{bmatrix}=-\begin{bmatrix}3x_1^{(k)}-\cos x_1^{(k)}-\sin x_2^{(k)}\\ 4x_2^{(k)}-\sin x_1^{(k)}-\cos x_2^{(k)}\end{bmatrix}\\ x^{(k+1)}=x^{(k)}+\Delta x^{(k)}\\ \qquad (k=0,1,\cdots)\end{cases}$$
选 $x^{(0)}=(1,1)^{\mathrm T}$，当 $k=4$ 时
$$\Delta x^{(4)}=\begin{bmatrix}-0.276\,244\times10^{-13}\\ -0.206\,813\times10^{-13}\end{bmatrix},\quad x^{(4)}=\begin{bmatrix}0.415\,169\,427\,139\\ 0.336\,791\,217\,025\end{bmatrix}$$
$$\|\Delta x^{(4)}\|_\infty/\|x^{(4)}\|_\infty=0.665\times10^{-13}<10^{-12}$$
故取 $x^*\approx x^{(4)}$。

Newton 法的优点是收敛快，一般都能达到平方收敛。但是，在许多情况下，Newton 法对初始向量 $x^{(0)}$ 的要求比较苛刻，往往要求 $x^{(0)}$ 很靠近解 x^*，Newton 法才收敛于 x^*。此外，Newton 法还须要求 $f_i(x)(i=1,2,\cdots,n)$ 的各个偏导数。为克服这些缺点，出现了许多种变形的 Newton 法。

4.2.4 离散 Newton 法

Newton 法(4.27)中的 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$，其元素$\dfrac{\partial f_i(\boldsymbol{x}^{(k)})}{\partial x_j}(i,j=1,2,\cdots,n)$如果用差商代替，则可得到新的迭代公式。令

$$\boldsymbol{h}^{(k)}=(h_1^{(k)},h_2^{(k)},\cdots,h_n^{(k)})^{\mathrm{T}},\quad h_j^{(k)}\neq 0\quad (j=1,2,\cdots,n)$$

$$\boldsymbol{J}(\boldsymbol{x}^{(k)},\boldsymbol{h}^{(k)})=\begin{bmatrix}\dfrac{f_1(\boldsymbol{x}^{(k)}+h_1^{(k)}\boldsymbol{e}_1)-f_1(\boldsymbol{x}^{(k)})}{h_1^{(k)}} & \cdots & \dfrac{f_1(\boldsymbol{x}^{(k)}+h_n^{(k)}\boldsymbol{e}_n)-f_1(\boldsymbol{x}^{(k)})}{h_n^{(k)}}\\ \vdots & & \vdots\\ \dfrac{f_n(\boldsymbol{x}^{(k)}+h_1^{(k)}\boldsymbol{e}_1)-f_n(\boldsymbol{x}^{(k)})}{h_1^{(k)}} & \cdots & \dfrac{f_n(\boldsymbol{x}^{(k)}+h_n^{(k)}\boldsymbol{e}_n)-f_n(\boldsymbol{x}^{(k)})}{h_n^{(k)}}\end{bmatrix}$$

其中 $\boldsymbol{e}_j$ 是第 j 个 n 维基本单位向量。迭代公式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\boldsymbol{J}(\boldsymbol{x}^{(k)},\boldsymbol{h}^{(k)})^{-1}\boldsymbol{F}(\boldsymbol{x}^{(k)})\quad (k=0,1,\cdots)\tag{4.28}$$

称为求解方程组(4.15)的离散 Newton 法。

为求方程组(4.15)的解 $\boldsymbol{x}^*\in\mathrm{int}(D)$，离散 Newton 法的算法如下：

(1) 在 $\boldsymbol{x}^*$ 附近选取 $\boldsymbol{x}^{(0)}\in D$，给定精度水平 $\varepsilon>0$ 和最大迭代次数 M。

(2) 对于 $k=0,1,\cdots,M$ 执行

① 选取 $\boldsymbol{h}^{(k)}=(h_1^{(k)},h_2^{(k)},\cdots,h_n^{(k)})^{\mathrm{T}},h_j^{(k)}\neq 0(j=1,2,\cdots,n)$。

② 计算 $\boldsymbol{F}(\boldsymbol{x}^{(k)})$和 $\boldsymbol{J}(\boldsymbol{x}^{(k)},\boldsymbol{h}^{(k)})$。

③ 求解关于 $\Delta\boldsymbol{x}^{(k)}$ 的线性方程组

$$\boldsymbol{J}(\boldsymbol{x}^{(k)},\boldsymbol{h}^{(k)})\Delta\boldsymbol{x}^{(k)}=-\boldsymbol{F}(\boldsymbol{x}^{(k)})$$

④ 若 $\|\Delta\boldsymbol{x}^{(k)}\|/\|\boldsymbol{x}^{(k)}\|\leqslant\varepsilon$，则取 $\boldsymbol{x}^*\approx\boldsymbol{x}^{(k)}$，并停止计算；否则转⑤。

⑤ 计算 $\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\Delta\boldsymbol{x}^{(k)}$。

⑥ 若 $k<M$，则继续；否则，输出 M 次迭代不成功的信息，并停止计算。

选取 $\boldsymbol{h}^{(k)}$ 除保证 $h_j^{(k)}\neq 0(j=1,2,\cdots,n)$外，还应使 $\boldsymbol{x}^{(k)}+h_j^{(k)}\boldsymbol{e}_j\in D(j=1,2,\cdots,n)$。有时可取 $h_j^{(k)}$ 为与 k 无关的非零常数 $h_j(j=1,2,\cdots,n)$。如果取 $h_j^{(k)}=c_j\|\boldsymbol{F}(\boldsymbol{x}^{(k)})\|$ (c_j 为非零常数；$j=1,2,\cdots,n$)，那么，此时的离散 Newton 法(4.28)称为 Newton - Steffensen(牛顿-斯蒂芬森)方法。在定理 4.15 Newton 法至少平方收敛的条件下，只要 c_j 选取恰当，Newton - Steffensen方法也收敛而且也是至少平方收敛。

习　题

1. 用对分法求下列方程的根，要求绝对误差限为 10^{-3}：

(1) $x^3-3x+1=0$ (只求最大的正根)；

(2) $x+\sin x=1$。

2. 用简单迭代法求下列方程的根，当满足$|x_k-x_{k-1}|/|x_k|\leqslant 10^{-6}$时结束迭代，并说明迭代收敛的理由：

(1) $\mathrm{e}^x+10x-2=0$；

(2) $x-\tan(x-1)=0$ (只求最小的正根)；

(3) $e^{-x}=\cos x$　(只求最小的正根)。

3. 证明下列迭代公式产生的序列$\{x_k\}$收敛,并判断有几阶收敛速度?

$$\begin{cases} x_0 \in [1.7,2] \\ x_{k+1} = 1+\dfrac{1}{x_k}+\dfrac{1}{x_k^2} \quad (k=0,1,\cdots) \end{cases}$$

4. 证明下列迭代公式产生的序列$\{x_k\}$收敛于$\sqrt{a}(a>0)$,并有三阶收敛速度:

$$x_{k+1} = \frac{x_k(x_k^2+3a)}{3x_k^2+a} \quad (k=0,1,\cdots)$$

其中 x_0 充分接近$\sqrt{a}$。

5. 试用 Steffensen 迭代法求方程

$$e^x + 10x - 2 = 0$$

的根,要求$|x_k - x_{k-1}|/|x_k| \leqslant 10^{-6}$。

6. 用 Newton 法求下列方程的根,要求$|x_k - x_{k-1}|/|x_k| \leqslant 10^{-6}$,并说明收敛的理由:

(1) $x+\sin x=1$;

(2) $x=5-2^x$;

(3) $x=\tan x$　(只求在 $x=100$ 附近的根)。

7. 函数 $f(x)$满足什么条件可使 Newton 法(4.7)具有三阶收敛速度? 其中 x_0 充分靠近方程 $f(x)=0$ 的根 s。

8. 设 s 是方程 $f(x)=0$ 的根,$f^{(4)}(x)$在 s 的某个邻域内连续,并且 $f'(s)\neq 0$。令

$$\varphi(x) = x - \frac{f(x)}{f'(x)} + h(x)\left[\frac{f(x)}{f'(x)}\right]^2$$

为使迭代法 $x_{k+1}=\varphi(x_k)(k=0,1,\cdots)$能够至少以三阶收敛速度收敛于 s,则应如何选取函数 $h(x)$?

9. 试分别用割线法和单点割线法求方程

$$x + \sin x = 1$$

的根,要求$|x_k - x_{k-1}|/|x_k| \leqslant 10^{-6}$。

10. 试用简单迭代法求方程组

$$\begin{cases} e^{-x_1} + 2x_2 = 1.97 \\ x_1 + e^{-2x_2} = 1.2 \end{cases}$$

在区域 $D=\{(x_1,x_2)\,|\,x_1\geqslant 0.5, x_2\geqslant 0.5\}$内的解,要求$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_\infty/\|\boldsymbol{x}^{(k)}\|_\infty\leqslant 0.0005$,并说明收敛的理由。

11. 试用 Newton 法求方程组

$$\begin{cases} x - \sin(x+y) = 1.2 \\ y + \cos(x+y) = 0.5 \end{cases}$$

在区域 $D=\{(x,y)\,|\,1\leqslant x\leqslant 1.5, 1\leqslant y\leqslant 1.5\}$内的解,要求$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_\infty/\|\boldsymbol{x}^{(k)}\|_\infty\leqslant 10^{-4}$,并说明收敛的理由。

第 5 章 插值与逼近

插值与逼近都是指用某个简单函数在满足一定条件下，在某个范围内近似代替另一个较为复杂或者解析表达式未给出的函数，以便于简化对后者的各种计算或揭示后者的某些性质。

5.1 代数插值

5.1.1 一元函数插值

定义 设有 $m+1$ 个互异的实数 $x_0,x_1,\cdots,x_m$ 和 $n+1$ 个实值函数 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$，其中 $n\leqslant m$。若向量组

$$\boldsymbol{\Phi}_k=(\varphi_k(x_0),\varphi_k(x_1),\cdots,\varphi_k(x_m))^{\mathrm{T}}\quad(k=0,1,\cdots,n)$$

线性无关，则称函数组 $\{\varphi_k(x)(k=0,1,\cdots,n)\}$ 在点集 $\{x_i(i=0,1,\cdots,m)\}$ 上线性无关；否则称为线性相关。

例如，函数组 $\{2+x,1-x,x+x^2\}$ 在点集 $\{1,2,3,4\}$ 上线性无关。

又如，函数组 $\{\sin x,\sin 2x,\sin 3x\}$ 在点集 $\left\{0,\frac{\pi}{3},\frac{2}{3}\pi,\pi\right\}$ 上线性相关。

给定 $n+1$ 个互异的实数 $x_0,x_1,\cdots,x_n$，实值函数 $f(x)$ 在包含 $x_0,x_1,\cdots,x_n$ 的某个区间 $[a,b]$ 内有定义。设函数组

$$\{\varphi_k(x)(k=0,1,\cdots,n)\}$$

是次数不高于 n 的多项式组，且在点集 $\{x_0,x_1,\cdots,x_n\}$ 上线性无关。

现在提出如下的问题：在次数不高于 n 的多项式集合

$$\mathscr{D}_n=\mathrm{Span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$$

中寻求多项式

$$p_n(x)=\sum_{k=0}^{n}c_k\varphi_k(x)\tag{5.1}$$

使其满足条件

$$p_n(x_i)=f(x_i)\quad(i=0,1,\cdots,n)\tag{5.2}$$

此问题称为一元函数的代数插值问题。$x_0,x_1,\cdots,x_n$ 称为插值节点；$f(x)$ 称为被插值函数；$\varphi_k(x)(k=0,1,\cdots,n)$ 称为插值基函数；条件(5.2)称为插值条件；满足插值条件(5.2)的多项式(5.1)称为 n 次插值多项式。

由于插值基函数组 $\{\varphi_k(x)(k=0,1,\cdots,n)\}$ 在点集 $\{x_0,x_1,\cdots,x_n\}$ 上线性无关，所以满足插值条件(5.2)的 n 次插值多项式 $p_n(x)$ 是存在且唯一的。

又由于插值基函数组限定为次数不高于 n 的多项式组，所以对于不同的插值基函数组，只要满足同一插值条件(5.2)，则所得的 n 次插值多项式也是唯一的。

今取插值基函数为

$$l_k(x)=\prod_{\substack{j=0\\j\neq k}}^{n}\frac{x-x_j}{x_k-x_j}\quad(k=0,1,\cdots,n)\tag{5.3}$$

显然，$l_k(x)(k=0,1,\cdots,n)$都是 n 次多项式，且具有下列性质

$$l_k(x_i)=\begin{cases}1, & i=k\\0, & i\neq k\end{cases}$$

因此，函数组$\{l_k(x)(k=0,1,\cdots,n)\}$必在点集$\{x_0,x_1,\cdots,x_n\}$上线性无关，并且

$$p_n(x)=\sum_{k=0}^{n}l_k(x)f(x_k)\tag{5.4}$$

或

$$p_n(x)=\sum_{k=0}^{n}\left(\prod_{\substack{j=0\\j\neq k}}^{n}\frac{x-x_j}{x_k-x_j}\right)f(x_k)\tag{5.5}$$

就是满足插值条件(5.2)的 n 次插值多项式。

形如式(5.5)的插值多项式称为 Lagrange(拉格朗日)插值多项式，而式(5.3)所表示的 n 次多项式称为 Lagrange 插值基函数，并称这种构造 n 次插值多项式的方法为 Lagrange 插值法。

定义　设实数 $x_0,x_1,\cdots,x_n$ 互异，称比值

$$f[x_0,x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0}$$

为 $f(x)$关于节点 x_0,x_1 的一阶差商。称比值

$$f[x_0,x_1,x_2]=\frac{f[x_0,x_2]-f[x_0,x_1]}{x_2-x_1}$$

为 $f(x)$关于节点 x_0,x_1,x_2 的二阶差商。一般地，设 $f(x)$的 $k-1$ 阶差商已定义，则称比值

$$f[x_0,x_1,\cdots,x_k]=\frac{f[x_0,x_1,\cdots,x_{k-2},x_k]-f[x_0,x_1,\cdots,x_{k-1}]}{x_k-x_{k-1}}$$

为 $f(x)$关于节点 $x_0,x_1,\cdots,x_k$ 的 k 阶差商$(k=2,3,\cdots,n)$。

特殊地，$f(x_i)$称为 $f(x)$关于 x_i 的零阶差商。

今取插值基函数为

$$\varphi_0(x)\equiv 1,\quad \varphi_k(x)=\prod_{j=0}^{k-1}(x-x_j)\quad(k=1,2,\cdots,n)$$

函数组$\{\varphi_k(x)(k=0,1,\cdots,n)\}$是次数不高于 n 的多项式组，且在点集$\{x_0,x_1,\cdots,x_n\}$上线性无关。因此，n 次插值多项式 $p_n(x)$可表示为

$$\begin{aligned}p_n(x)=&c_0+c_1(x-x_0)+c_2(x-x_0)(x-x_1)+\cdots+\\&c_n(x-x_0)(x-x_1)\cdots(x-x_{n-1})\end{aligned}\tag{5.6}$$

令 $p_n(x)$满足插值条件(5.2)，得到关于系数 $c_0,c_1,\cdots,c_n$ 的下三角形线性方程组

$$\begin{aligned}&p_n(x_0)=c_0=f(x_0)\\&p_n(x_1)=c_0+c_1(x_1-x_0)=f(x_1)\\&p_n(x_2)=c_0+c_1(x_2-x_0)+c_2(x_2-x_0)(x_2-x_1)=f(x_2)\\&\quad\vdots\\&p_n(x_n)=c_0+c_1(x_n-x_0)+\cdots+c_n(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})=f(x_n)\end{aligned}$$

由此方程组可递推地求出系数 $c_0,c_1,\cdots,c_n$，并可用差商表示，结果如下：

$$c_0 = f(x_0)$$

$$c_1 = \frac{f(x_1)-f(x_0)}{x_1-x_0} = f[x_0,x_1]$$

$$c_2 = \frac{f(x_2)-f(x_0)-c_1(x_2-x_0)}{(x_2-x_0)(x_2-x_1)} = f[x_0,x_1,x_2]$$

使用数学归纳法可以证明：

$$c_k = f[x_0,x_1,\cdots,x_k] \quad (k=1,2,\cdots,n)$$

把所得到的 c_k 代入式(5.6)就得到另一种形式的 n 次插值多项式

$$\begin{aligned} p_n(x) = {} & f(x_0)+f[x_0,x_1](x-x_0)+ \\ & f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots+ \\ & f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}) \end{aligned} \tag{5.7}$$

形如式(5.7)的 n 次插值多项式称为 Newton 插值多项式，而这种插值法称为 Newton 插值法。

根据满足插值条件(5.2)的 n 次插值多项式的唯一性，任意交换节点 $x_0,x_1,\cdots,x_n$ 的次序所得到的各个 n 次 Newton 插值多项式都是同一个 n 次多项式。特别是，由于 x^n 项的系数均相同，因而可知，$f(x)$关于节点 $x_0,x_1,\cdots,x_n$ 的 n 阶差商与节点的排列次序无关。于是有

$$\begin{aligned} f[x_0,x_1,\cdots,x_n] &= f[x_1,x_2,\cdots,x_n,x_0] = \\ & \frac{f[x_1,x_2,\cdots,x_{n-1},x_0]-f[x_1,x_2,\cdots,x_n]}{x_0-x_n} = \\ & \frac{f[x_1,x_2,\cdots,x_n]-f[x_0,x_1,\cdots,x_{n-1}]}{x_n-x_0} \end{aligned}$$

由此可知，下列等式成立，

$$f[x_i,x_{i+1},\cdots,x_{i+k}] = \frac{f[x_{i+1},x_{i+2},\cdots,x_{i+k}]-f[x_i,x_{i+1},\cdots,x_{i+k-1}]}{x_{i+k}-x_i}$$

根据差商的这个性质，就可由数表$(x_i,f(x_i))(i=0,1,\cdots,n)$造出差商表，逐列计算 $f(x)$的各阶差商，从而得出 Newton 插值多项式(5.7)中的各个系数 $c_k=f[x_0,x_1,\cdots,x_k](k=0,1,\cdots,n)$。四个节点的差商表见表 5-1。

表 5-1 差商表

x	$f(x)$	一阶差商	二阶差商	三阶差商
x_0	$f(x_0)$			
x_1	$f(x_1)$	$f[x_0,x_1]$		
x_2	$f(x_2)$	$f[x_1,x_2]$	$f[x_0,x_1,x_2]$	
x_3	$f(x_3)$	$f[x_2,x_3]$	$f[x_1,x_2,x_3]$	$f[x_0,x_1,x_2,x_3]$

由于满足条件(5.2)的 n 次插值多项式是唯一的，所以，Lagrange插值多项式(5.5)和 Newton 插值多项式(5.7)是同一个n 次多项式。

用 n 次插值多项式 $p_n(x)$近似替代被插值函数 $f(x)$，得近似等式

$$f(x) \approx p_n(x) \tag{5.8}$$

式(5.8)称为插值公式，并称误差

$$R_n(x) = f(x) - p_n(x)$$

为插值公式(5.8)的余项或截断误差。

用记号 I_x 表示包含实数 $x, x_0, x_1, \cdots, x_n$ 的最小闭区间，记号 $\tilde{I}_x$ 表示含于 I_x 中的最大开区间，又记

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$$

定理 5.1　设 $x_0, x_1, \cdots, x_n$ 是互异的实数，对于给定的实数 x，实值函数 $f(t)$ 在区间 I_x 上具有 $n+1$ 阶导数，则插值公式(5.8)的余项为

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x) \tag{5.9}$$

其中 $\xi \in \tilde{I}_x$ 且依赖于 x。

证　当给定的 x 恰是某个节点 x_i 时，式(5.9)两边都为零，定理的结论显然成立。今设给定的 x 异于所有的节点。构造辅助函数

$$g(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\omega_{n+1}(x)}\omega_{n+1}(t)$$

因 $f(t)$，$p_n(t)$ 和 $\omega_{n+1}(t)$ 都在 I_x 上 $n+1$ 次可微，故函数 $g(t)$ 也如此。显然，函数 $g(t)$ 有 $n+2$ 个互异的零点 $x, x_0, x_1, \cdots, x_n$。由 Rolle(罗尔)定理可知，$g'(t)$ 在区间 $\tilde{I}_x$ 内至少有 $n+1$ 个互异的零点。再对函数 $g'(t)$ 使用 Rolle 定理，可知在 $\tilde{I}_x$ 内至少有 n 个互异的点使 $g''(t)=0$。如此反复使用 Rolle 定理，最后可知至少存在一点 $\xi \in \tilde{I}_x$，使得

$$g^{(n+1)}(\xi) = 0$$

ξ 显然与所给的 x 有关。由于

$$g^{(n+1)}(t) = f^{(n+1)}(t) - \frac{f(x) - p_n(x)}{\omega_{n+1}(x)}(n+1)!$$

因而有

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)$$

其中 $\xi \in \tilde{I}_x$ 且依赖于 x。

证毕。

若已知常数 M_{n+1} 满足

$$\max_{t \in \tilde{I}_x} |f^{(n+1)}(t)| \leqslant M_{n+1}$$

则余项的估计式是

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|$$

从余项表达式(5.9)可看出，当 $f(x)$ 是次数不高于 n 的多项式时，无论 $n+1$ 个互异的插值节点 $x_0, x_1, \cdots, x_n$ 如何选取，$f(x)$ 的 n 次插值多项式 $p_n(x)$ 就是 $f(x)$ 本身。特别是，由 $f(x) \equiv 1$ 的 n 次 Lagrange 插值多项式可推知，任意的 $n+1$ 个互异的 $x_0, x_1, \cdots, x_n$ 形成的 Lagrange插值基函数(5.3)具有下列性质

$$\sum_{k=0}^{n} l_k(x) \equiv 1$$

由 Newton 插值多项式的构成方法可知，若 x 是异于 $x_0, x_1, \cdots, x_n$ 的实数，则函数 $f(t)$ 的

以 $x_0,x_1,\cdots,x_n,x$ 为插值节点的 $n+1$ 次 Newton 插值多项式为

$$p_{n+1}(t)=p_n(t)+f[x_0,x_1,\cdots,x_n,x](t-x_0)(t-x_1)\cdots(t-x_n)$$

因而有

$$f(x)=p_{n+1}(x)=p_n(x)+f[x_0,x_1,\cdots,x_n,x](x-x_0)(x-x_1)\cdots(x-x_n)$$

其中 $p_n(x)$是 n 次 Newton 插值多项式(5.7)。因此,插值公式(5.8)的余项也可表示为

$$R_n(x)=f[x_0,x_1,\cdots,x_n,x](x-x_0)(x-x_1)\cdots(x-x_n)$$

由此得到 $f(x)$的 $n+1$ 阶差商与 $f(x)$的 $n+1$ 阶导数的关系

$$f[x_0,x_1,\cdots,x_n,x]=\frac{f^{(n+1)}(\xi)}{(n+1)!},\quad \xi\in\tilde{I}_x$$

例 1 试由 $f(x)=2^x$ 的函数表

x	-1	0	1
$f(x)$	0.5	1	2

建立二次插值多项式 $p_2(x)$,用以计算 $2^{0.3}$的近似值,并估计截断误差。

解 使用 $n=2$ 的 Lagrange 插值多项式(5.5),得

$$p_2(x)=\frac{(x-0)(x-1)}{(-1-0)(-1-1)}0.5+\frac{(x+1)(x-1)}{(0+1)(0-1)}1+\frac{(x+1)(x-0)}{(1+1)(1-0)}2=0.25x^2+0.75x+1$$

也可使用 $n=2$ 的 Newton 插值多项式(5.7)。由差商表 5-2 可得

表 5-2 $f(x)=2^x$ 的差商表

x	$f(x)$	一阶差商	二阶差商
-1	0.5		
0	1	$f[-1,0]=0.5$	$f[-1,0,1]=0.25$
1	2	$f[0,1]=1$	

$$p_2(x)=f(-1)+f[-1,0](x+1)+f[-1,0,1](x+1)(x-0)=0.5+0.5(x+1)+0.25(x+1)x=0.25x^2+0.75x+1$$

$$2^{0.3}\approx p_2(0.3)=1.247\,5$$

因 $f'''(x)=2^x(\ln 2)^3$, $\max\limits_{-1\leqslant x\leqslant 1}|f'''(x)|=2(\ln 2)^3=0.666\,0$,故有

$$|2^{0.3}-p_2(0.3)|\leqslant\frac{0.666\,0}{3!}|(0.3+1)(0.3-0)(0.3-1)|=0.030\,30$$

可见,用 $p_2(0.3)=1.247\,5$ 作为 $2^{0.3}$的近似值,可保证有两位有效数字。

利用插值多项式 $p_n(x)$计算被插值函数 $f(x)$在区间$[a,b]$上任一点 x 处的近似值,总希望截断误差的绝对值$|R_n(x)|$更小一些。影响$|R_n(x)|$大小的因素主要是插值多项式的次数 n和插值节点的选择。理论上和计算实践都说明高次插值不可取。Runge(龙格)给出了一个例子,$f(x)=\frac{1}{1+x^2}$在区间$[-5,5]$上使用 $n+1$ 个等距节点 $x_i=-5+\frac{10}{n}i\,(i=0,1,\cdots,n)$作 n 次插值。采用Lagrange插值多项式的形式,得 n 次插值多项式

$$p_n(x)=\sum_{k=0}^{n}\left[\prod_{\substack{j=0\\j\neq k}}^{n}\frac{x+5-\frac{10}{n}j}{\frac{10}{n}(k-j)}\right]\frac{1}{1+\left(-5+\frac{10}{n}k\right)^2}$$

取 $x=3.6$ 和 $x=4.7$，分别作 $n=10$ 次和 $n=20$ 次插值，结果为

$$p_{10}(3.6)=-0.2546027,\quad p_{10}(4.7)=1.95895$$

$$p_{20}(3.6)=-0.1143479,\quad p_{20}(4.7)=-28.6626$$

其插值结果与

$$f(3.6)=0.0716332378,\quad f(4.7)=0.04330879$$

相差甚远。Runge 已经证明，当 $n\to\infty$ 时，$p_n(x)$ 不在整个区间 $[-5,5]$ 上收敛于 $f(x)$。

在实际计算中，通常采用分段低次插值。

设节点等距，$x_i=x_0+ih\ (h>0)$，$f(x_i)\ (i=0,1,\cdots,n)$ 已知。对给定的 x，要采用 m 次插值 $(1\leqslant m\leqslant 3)$ 计算 $f(x)$ 的近似值。为使 $|R_m(x)|$ 中的因子 $|\omega_{m+1}(x)|$ 尽量小，应选择最靠近 x 的 $m+1$ 个节点作为插值节点。例如，采用二次插值，应按以下方法选择三个插值节点：若 $x\leqslant x_1+\frac{h}{2}$，则选 x_0,x_1,x_2；若 $x>x_{n-1}-\frac{h}{2}$，则选 x_{n-2},x_{n-1},x_n；若 $x_j-\frac{h}{2}<x\leqslant x_j+\frac{h}{2}\ (2\leqslant j\leqslant n-2)$，则选 x_{j-1},x_j,x_{j+1}。

在区间 $[a,b]$ 上分段低次插值所得的插值函数在区间 $[a,b]$ 上一般是不连续的，或者虽然插值函数连续，但其导数不连续。因此，这种分段低次插值不适用于光滑性要求高的外形设计。

5.1.2 二元函数插值

设 $a\leqslant x_0<x_1<\cdots<x_{n-1}<x_n\leqslant b,c\leqslant y_0<y_1<\cdots<y_{m-1}<y_m\leqslant d$，函数 $f(x,y)$ 是定义在矩形域 $D=\{a\leqslant x\leqslant b,c\leqslant y\leqslant d\}$ 上的实值函数。取点集 $\{(x_i,y_j)(i=0,1,\cdots,n;j=0,1,\cdots,m)\}$ 为插值节点，取在插值节点集上线性无关的函数组 $\{\varphi_{kr}(x,y)(k=0,1,\cdots,n;r=0,1,\cdots,m)\}$ 为插值基函数组，其中 $\varphi_{kr}(x,y)$ 是 x 不高于 n 次、y 不高于 m 次的二元多项式。在集合

$$\mathscr{D}=\mathrm{Span}\{\varphi_{00},\cdots,\varphi_{0m},\cdots,\varphi_{n0},\cdots,\varphi_{nm}\}$$

中寻求二元插值多项式

$$p_{nm}(x,y)=\sum_{k=0}^{n}\sum_{r=0}^{m}c_{kr}\varphi_{kr}(x,y)\tag{5.10}$$

使其满足插值条件

$$p_{nm}(x_i,y_j)=f(x_i,y_j)\quad(i=0,1,\cdots,n;j=0,1,\cdots,m)\tag{5.11}$$

此问题就是二元函数的代数插值问题。

由于插值基函数组 $\{\varphi_{kr}(x,y)(k=0,1,\cdots,n;r=0,1,\cdots,m)\}$ 在插值节点集上线性无关，所以，满足插值条件(5.11)的二元插值多项式(5.10)是存在且唯一的。

现取插值基函数

$$\varphi_{kr}(x,y)=l_k(x)\tilde{l}_r(y)\quad(k=0,1,\cdots,n;r=0,1,\cdots,m)$$

其中

$$l_k(x)=\prod_{\substack{t=0\\t\neq k}}^{n}\frac{x-x_t}{x_k-x_t},\quad \tilde{l}_r(y)=\prod_{\substack{t=0\\t\neq r}}^{m}\frac{y-y_t}{y_r-y_t}\tag{5.12}$$

这样的 $\varphi_{kr}(x,y)$是 x 为 n 次、y 为 m 次的二元多项式，且满足

$$\varphi_{kr}(x_i,y_j)=\begin{cases}1, & (i,j)=(k,r)\\ 0, & (i,j)\neq(k,r)\end{cases}$$

因而基函数组$\{\varphi_{kr}(x,y)\}$必在点集$\{(x_i,y_j)\}$上线性无关，并且可知满足插值条件(5.11)的二元插值多项式为

$$p_{nm}(x,y)=\sum_{k=0}^{n}\sum_{r=0}^{m}l_k(x)\tilde{l}_r(y)f(x_k,y_r) \tag{5.13}$$

其中 $l_k(x)$和 $\tilde{l}_r(y)$由式(5.12)给出。式(5.13)称为 Lagrange 形式的插值曲面。

近似等式

$$f(x,y)\approx p_{nm}(x,y) \tag{5.14}$$

称为二元函数的 Lagrange 插值公式，式中 $p_{nm}(x,y)$由式(5.13)给出。称

$$R_{nm}(x,y)=f(x,y)-p_{nm}(x,y)$$

为插值公式(5.14)的余项或截断误差。利用一元函数插值公式的余项(5.9)可推出 $R_{nm}(x,y)$的表达式。

设 $f(x,y)$在区域 D 上对 x 有 $n+1$ 阶偏导数、对 y 有 $m+1$ 阶偏导数，并设$(x,y)\in D$，则由式(5.9)可得

$$\begin{aligned}R_{nm}(x,y)&=f(x,y)-\sum_{k=0}^{n}l_k(x)f(x_k,y)+\\&\sum_{k=0}^{n}l_k(x)f(x_k,y)-\sum_{r=0}^{m}\tilde{l}_r(y)\sum_{k=0}^{n}l_k(x)f(x_k,y_r)=\\&\frac{\omega_{n+1}(x)}{(n+1)!}f_{x^{n+1}}^{(n+1)}(\xi,y)+\frac{\omega_{m+1}(y)}{(m+1)!}\sum_{k=0}^{n}l_k(x)f_{y^{m+1}}^{(m+1)}(x_k,\eta)\end{aligned} \tag{5.15}$$

其中 $\xi\in(a,b)$，$\eta\in(c,d)$且都与(x,y)有关，而 $\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$，$\omega_{m+1}(y)=(y-y_0)(y-y_1)\cdots(y-y_m)$。同理又可得

$$R_{nm}(x,y)=\frac{\omega_{m+1}(y)}{(m+1)!}f_{y^{m+1}}^{(m+1)}(x,\eta)+\frac{\omega_{n+1}(x)}{(n+1)!}\sum_{r=0}^{m}\tilde{l}_r(y)f_{x^{n+1}}^{(n+1)}(\xi,y_r) \tag{5.16}$$

其中 $\eta\in(c,d)$，$\xi\in(a,b)$且都与(x,y)有关。

由余项表达式(5.15)或(5.16)可知，若被插值函数 $f(x,y)$是 x 为 n 次、y 为 m 次的二元多项式

$$f(x,y)=\sum_{k=0}^{n}\sum_{r=0}^{m}c_{kr}x^ky^r$$

则 $f(x,y)$的二元插值多项式(5.13)就是 $f(x,y)$本身。

若已知常数 M 和 N 满足

$$\max_{(x,y)\in D}\left|f_{x^{n+1}}^{(n+1)}(x,y)\right|\leqslant M,\quad \max_{(x,y)\in D}\left|f_{y^{m+1}}^{(m+1)}(x,y)\right|\leqslant N$$

则由式(5.15)可得余项的估计式

$$|R_{nm}(x,y)|\leqslant\frac{|\omega_{n+1}(x)|}{(n+1)!}M+\frac{|\omega_{m+1}(y)|}{(m+1)!}N\sum_{k=0}^{n}|l_k(x)|$$

例 2 试利用 $f(x,y)$的函数表

f \ y \ x	-1	0	1
0.5	0.25	0.5	1
1	0.43	0.87	1.73

建立 x 为二次、y 为一次的二元插值多项式 $p_{21}(x,y)$，用以计算 $f(0.3,0.8)$ 的近似值。

解 由 $n=2,m=1$ 的二元插值多项式(5.13)可得

$$p_{21}(x,y)=\frac{1}{2}x(x-1)\left(\frac{y-1}{-0.5}0.25+\frac{y-0.5}{0.5}0.43\right)-$$

$$(x+1)(x-1)\left(\frac{y-1}{-0.5}0.5+\frac{y-0.5}{0.5}0.87\right)+$$

$$\frac{1}{2}x(x+1)\left(\frac{y-1}{-0.5}+\frac{y-0.5}{0.5}1.73\right)=$$

$$0.17x^2y+0.55xy+0.04x^2+0.1x+0.74y+0.13$$

$$f(0.3,0.8)\approx p_{21}(0.3,0.8)=0.899\ 84$$

与一元函数的情形相类似，二元函数的代数插值也不宜使用高次插值，而应采取分片低次插值。常用的是分片双一次或双二次或对一个变量一次对另一个变量二次插值。设

$$x_i=x_0+ih \quad (i=0,1,\cdots,n)$$

$$y_j=y_0+j\tau \quad (j=0,1,\cdots,m)$$

$h>0,\tau>0$，则当 (x,y) 给定之后，根据 x 选择 x_i 和根据 y 选择 y_j 的原则与一元函数情形相同。例如，要采取双二次插值，且 (x,y) 满足

$$x_i-\frac{h}{2}<x\leqslant x_i+\frac{h}{2},\quad 2\leqslant i\leqslant n-2$$

$$y_j-\frac{\tau}{2}<y\leqslant y_j+\frac{\tau}{2},\quad 2\leqslant j\leqslant m-2$$

则应选择 $(x_k,y_r)(k=i-1,i,i+1;r=j-1,j,j+1)$ 为插值节点，相应的插值多项式为

$$p_{22}(x,y)=\sum_{k=i-1}^{i+1}\sum_{r=j-1}^{j-1}l_k(x)\tilde{l}_r(y)f(x_k,y_r) \tag{5.17}$$

其中

$$l_k(x)=\prod_{\substack{t=i-1\\t\neq k}}^{i+1}\frac{x-x_t}{x_k-x_t} \quad (k=i-1,i,i+1)$$

$$\tilde{l}_r(y)=\prod_{\substack{t=j-1\\t\neq r}}^{j-1}\frac{y-y_t}{y_r-y_t} \quad (r=j-1,j,j+1)$$

如果 $x\leqslant x_1+\frac{h}{2}$ 或 $x>x_{n-1}-\frac{h}{2}$，则在式(5.17)中取 $i=1$ 或 $i=n-1$；如果 $y\leqslant y_1+\frac{\tau}{2}$ 或 $y>y_{m-1}-\frac{\tau}{2}$，则在式(5.17)中取 $j=1$ 或 $j=m-1$。

5.2 Hermite 插值

前面所讨论的代数插值问题只要求插值多项式 $p_n(x)$ 满足插值条件(5.2)，这些条件仅对节点处的函数值作了约束，因而所得的插值多项式不能全面反映被插值函数 $f(x)$ 的性态。如

果插值条件再增加对节点处导数的限制，则所构造的多项式一定能更好地逼近函数 $f(x)$。

给定 $n+1$ 个互异节点 $x_0,x_1,\cdots,x_n$，并已知函数值 $y_i=f(x_i)(i=0,1,\cdots,n)$ 以及导数值 $y'_{i_k}=f'(x_{i_k})(k=0,1,\cdots,m)$，其中，标号 $i_0,i_1,\cdots,i_m$ 互异，$0\leqslant i_k\leqslant n,m\leqslant n$。问题是：在次数不高于 $m+n+1$ 的多项式集合 $\mathscr{D}_{m+n+1}$ 中求一多项式

$$H_{m+n+1}(x)=\sum_{j=0}^{m+n+1}a_jx^j \tag{5.18}$$

使其满足条件

$$\begin{cases} H_{m+n+1}(x_i)=y_i \quad (i=0,1,\cdots,n) & (5.19)\\ H'_{m+n+1}(x_{i_k})=y'_{i_k} \quad (k=0,1,\cdots,m) & (5.20)\end{cases}$$

此问题称为 Hermite(埃尔米特)插值问题，也称为带导数条件的代数插值问题。满足条件(5.19)、(5.20)的多项式(5.18)称为 Hermite 插值多项式。

定理 5.2 设 $n+1$ 个节点 $x_0,x_1,\cdots,x_n$ 互异，则在多项式集合 $\mathscr{D}_{m+n+1}$ 中，唯一地存在多项式(5.18)，满足条件(5.19)、(5.20)。

证

先证存在性。令

$$H_{m+n+1}(x)=p_n(x)+q_m(x)\omega_{n+1}(x) \tag{5.21}$$

其中 $p_n(x)$是满足插值条件(5.2)的 n 次插值多项式，$q_m(x)=\sum_{k=0}^{m}a_kx^k$。显然，$H_{m+n+1}(x)\in\mathscr{D}_{m+n+1}$，并且，对任意的 $q_m(x)$，多项式(5.21)已满足条件(5.19)。由条件(5.20)得

$$H'_{m+n+1}(x_{i_k})=p'_n(x_{i_k})+q_m(x_{i_k})\prod_{\substack{j=0\\ j\neq i_k}}^{n}(x_{i_k}-x_j)=y'_{i_k}$$

因而有

$$q_m(x_{i_k})=[y'_{i_k}-p'_n(x_{i_k})]\Big/\prod_{\substack{j=0\\ j\neq i_k}}^{n}(x_{i_k}-x_j) \quad (k=0,1,\cdots,m) \tag{5.22}$$

式(5.22)是一个以 $a_0,a_1,\cdots,a_m$ 为未知数的线性方程组。因函数组$\{1,x,\cdots,x^m\}$在点集$\{x_{i_k}(k=0,1,\cdots,m)\}$上线性无关，故线性方程组(5.22)有唯一解 $a_0,a_1,\cdots,a_m$。因而多项式(5.21)被确定，它满足条件(5.19)、(5.20)。

再证唯一性。设存在两个多项式 $H(x),\tilde{H}(x)\in\mathscr{D}_{m+n+1}$，它们都满足条件(5.19)和(5.20)。记 $r(x)=H(x)-\tilde{H}(x)$，则有

$$r(x_i)=0 \quad (i=0,1,\cdots,n)$$

$$r'(x_{i_k})=0 \quad (k=0,1,\cdots,m)$$

因而 $r(x)$有 $m+1$ 个二重零点和 $n-m$ 个单零点。由于 $2(m+1)+n-m=m+n+2$，而 $r(x)\in\mathscr{D}_{m+n+1}$，所以只能 $r(x)\equiv0$，即$H(x)\equiv\tilde{H}(x)$。

证毕。

定理 5.3 设 $x_0,x_1,\cdots,x_n$ 是互异的实数，对于给定的 x，实值函数 $f(t)$在区间 I_x 上具有 $m+n+2$阶导数，$H_{m+n+1}(x)$是满足条件(5.19)、(5.20)的 Hermite 插值多项式，则用 $H_{m+n+1}(x)$近似代替 $f(x)$的余项为

$$R(x)=f(x)-H_{m+n+1}(x)=\frac{f^{(m+n+2)}(\xi)}{(m+n+2)!}\omega_{n+1}(x)\prod_{k=0}^{m}(x-x_{i_k}) \tag{5.23}$$

其中 $\xi\in\tilde{I}_x$ 且依赖于 x。

证　当 x 恰是某个节点 x_i 时，式(5.23)两边都为零，定理结论成立，ξ 可为 $\tilde{I}_x$ 内任一点。今设 x 异于所有的节点。构造辅助函数

$$g(t)=f(t)-H_{m+n+1}(t)-\frac{f(x)-H_{m+n+1}(x)}{\omega_{n+1}(x)\prod_{k=0}^{m}(x-x_{i_k})}\omega_{n+1}(t)\prod_{k=0}^{m}(t-x_{i_k})$$

易知，$x,x_0,x_1,\cdots,x_n$ 是 $g(t)$ 的零点。根据 Rolle 定理，$g'(t)$ 在区间 $\tilde{I}_x$ 内至少有 $n+1$ 个异于 $x,x_0,x_1,\cdots,x_n$ 的零点；又 $x_{i_k}(k=0,1,\cdots,m)$ 也是 $g'(t)$ 的零点，故 $g'(t)$ 在区间 $\tilde{I}_x$ 内至少有 $m+n+2$ 个互异零点。继续使用 Rolle 定理，可知 $g^{(m+n+2)}(t)$ 在 $\tilde{I}_x$ 内至少有一个零点 ξ，即

$$g^{(m+n+2)}(\xi)=0$$

其中 $\xi\in\tilde{I}_x$ 且依赖于 x。但

$$g^{(m+n+2)}(t)=f^{(m+n+2)}(t)-\frac{f(x)-H_{m+n+1}(x)}{\omega_{n+1}(x)\prod_{k=0}^{m}(x-x_{i_k})}(m+n+2)!$$

用 $t=\xi$ 代入上式，并解出 $f(x)-H_{m+n+1}(x)$，即得式(5.23)。

证毕。

满足条件(5.19)、(5.20)的 Hermite 插值多项式 $H_{m+n+1}(x)$，在几何上表示曲线 $y=H_{m+n+1}(x)$ 不仅要通过点 $(x_i,y_i)(i=0,1,\cdots,n)$，而且在点 $(x_{i_k},y_{i_k})(k=0,1,\cdots,m)$ 处，曲线 $y=H_{m+n+1}(x)$ 与曲线 $y=f(x)$ 有共同的切线。

例 3　给定数表

x	-1	0	1	2
$f(x)$	10	14	16	15
$f'(x)$	1		0.1	

求次数不高于 5 的多项式 $H_5(x)$，使其满足条件

$$\begin{cases}H_5(x_i)=f(x_i) & (i=0,1,2,3)\\ H'_5(x_i)=f'(x_i) & (i=0,2)\end{cases}$$

其中 $x_i=-1+i(i=0,1,2,3)$。

解　先建立满足条件

$$p_3(x_i)=f(x_i)\quad(i=0,1,2,3)$$

的三次插值多项式 $p_3(x)$，现采用 Newton 插值多项式

$$\begin{aligned}p_3(x)=&f(x_0)+f[x_0,x_1](x-x_0)+\\&f[x_0,x_1,x_2](x-x_0)(x-x_1)+\\&f[x_0,x_1,x_2,x_3](x-x_0)(x-x_1)(x-x_2)=\\&10+4(x+1)-(x+1)x-\frac{1}{6}(x+1)x(x-1)=\\&14+\frac{19}{6}x-x^2-\frac{1}{6}x^3\end{aligned}$$

再设

$$H_5(x) = p_3(x) + (ax+b)(x+1)x(x-1)(x-2)$$

由

$$\begin{cases} H'_5(-1) = p'_3(-1) + (-a+b)(-6) = 1 \\ H'_5(1) = p'_3(1) + (a+b)(-2) = 0.1 \end{cases}$$

得

$$\begin{cases} -a+b = \dfrac{11}{18} \\ a+b = \dfrac{17}{60} \end{cases}$$

解出

$$a = -\frac{59}{360}, \quad b = \frac{161}{360}$$

故所求的插值多项式为

$$H_5(x) = 14 + \frac{19}{6}x - x^2 - \frac{1}{6}x^3 + \frac{1}{360}(161-59x)x(x^2-1)(x-2)$$

5.3 样条插值

5.3.1 样条函数

一元函数的样条插值是在区间$[a,b]$上用分段低次多项式作为插值函数，且能满足对光滑性的要求。它除了要求给出$[a,b]$内各个节点处的被插值函数值外，还须提供两个边界节点处的导数信息。

样条插值所用的分段低次多项式就是样条函数。

定义 记

$$x_+^k = \begin{cases} x^k, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \quad (k = 1,2,\cdots)$$

称 $x_+^k\ (k=1,2,\cdots)$为 k 次半截单项式，并规定

$$x_+^0 = \begin{cases} 1, & x > 0 \\ \dfrac{1}{2}, & x = 0 \\ 0, & x < 0 \end{cases}$$

由定义可知，$x_+^k\ (k=1,2,\cdots)$在区间$(-\infty,\infty)$上有 $k-1$ 阶连续导数，且当 $k\geqslant 2$ 时有

$$(x_+^k)^{(r)} = k(k-1)\cdots(k-r+1)x_+^{k-r} \quad (r = 1,2,\cdots,k-1)$$

但 x_+^k 在 $x=0$ 处 k 阶导数不存在。

对任何实数 a，有

$$(x-a)_+^k = \begin{cases} (x-a)^k, & x \geqslant a \\ 0, & x < a \end{cases} \quad (k = 1,2,\cdots)$$

定义　对于区间$[a,b]$上的一个分划

$$\pi:\quad a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

如果函数$s(x)$满足条件

(1) $s(x)$在每个子区间$[x_i, x_{i+1}]$ $(i=0,1,\cdots,n-1)$上是次数不高于k的多项式；

(2) $s(x)$在区间(a,b)上有$k-1$阶连续导数。

则称$s(x)$是定义在$[a,b]$上对应于分划π的k次多项式样条函数(简称k次样条)。$x_0,x_1,\cdots,x_n$称为样条节点,其中$x_1,x_2,\cdots,x_{n-1}$称为内节点,x_0和x_n称为边界节点。

定义中的区间$[a,b]$可以是$(-\infty,\infty)$,这时看做$x_0=-\infty, x_n=\infty$。

由于k次样条函数$s(x)$在每个子区间上都是具有$k+1$个系数的k次多项式,分划π共有n个子区间,故对应于分划π的k次样条$s(x)$共有$n(k+1)$个待定参数。由定义中的条件(2)提供了$k(n-1)$个约束条件

$$s^{(r)}(x_i^-) = s^{(r)}(x_i^+)\quad (i = 1,2,\cdots,n-1; r = 0,1,\cdots,k-1)$$

即$s(x)$及其1至$k-1$阶导数在内节点处的左右极限应相等。由于$n(k+1)-k(n-1)=n+k$,所以,对应于分划π的k次样条组成了一个函数集合,而且是一个线性空间,记为$\mathscr{D}_{k,\pi}$,它的维数不超过$n+k$。

根据样条函数定义,易知

$$x^j \in \mathscr{D}_{k,\pi}\quad (j = 0,1,\cdots,k)$$

$$(x-x_j)_+^k \in \mathscr{D}_{k,\pi}\quad (j = 1,2,\cdots,n-1)$$

定义　设$\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$在区间$[a,b]$上连续,如果

$$a_0\varphi_0(x) + a_1\varphi_1(x) + \cdots + a_n\varphi_n(x) \equiv 0,\quad a \leqslant x \leqslant b$$

当且仅当$a_0=a_1=\cdots=a_n=0$才成立,就称函数组$\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$在区间$[a,b]$上线性无关;否则称为线性相关。若函数系$\{\varphi_j(x)(j=0,1,\cdots)\}$中任何有限个函数在$[a,b]$上线性无关,则称它为在区间$[a,b]$上线性无关的函数系。

空间$\mathscr{D}_{k,\pi}$中的函数组

$$\{x^j(j = 0,1,\cdots,k),(x-x_j)_+^k\ (j = 1,2,\cdots,n-1)\}\tag{5.24}$$

就是在$[a,b]$上线性无关的函数组。事实上,若

$$\sum_{j=0}^{k} a_j x^j + \sum_{j=1}^{n-1} b_j (x-x_j)_+^k \equiv 0,\quad a \leqslant x \leqslant b$$

那么,令$x<x_1$,推得$\sum_{j=0}^{k} a_j x^j \equiv 0$,由此可知$a_j=0(j=0,1,\cdots,k)$;令$x_1<x<x_2$可推出$b_1(x-x_1)^k\equiv 0$,因而$b_1=0$。设已推出$b_1=b_2=\cdots=b_{i-1}=0$,令$x_i<x<x_{i+1}$可推出$b_i(x-x_i)^k\equiv 0$,因而$b_i=0$。当$i=n-1$时,就得到$b_j=0(j=1,2,\cdots,n-1)$。

因此,线性空间$\mathscr{D}_{k,\pi}$的维数就是$n+k$,并可用函数组(5.24)作基底,即

$$\mathscr{D}_{k,\pi} = \mathrm{Span}\{1,x,\cdots,x^k,(x-x_1)_+^k,\cdots,(x-x_{n-1})_+^k\}$$

于是定义在$[a,b]$上对应于分划π的k次样条函数总可表示为

$$s(x) = \sum_{j=0}^{k} a_j x^j + \frac{1}{k!}\sum_{j=1}^{n-1} c_j (x-x_j)_+^k,\quad a \leqslant x \leqslant b\tag{5.25}$$

其中$a_j(j=0,1,\cdots,k)$,$c_j(j=1,2,\cdots,n-1)$为$n+k$个自由参数。要想在空间$\mathscr{D}_{k,\pi}$中确定一个$s(x)$需要$n+k$个条件。

由于$(x-x_j)_+^k$在x_j处不存在k阶导数，所以，一般情况下，k次样条在其内节点处不存在k阶导数。

记

$$\Omega_k(x)=\frac{1}{k!}\sum_{j=0}^{k+1}(-1)^j\binom{k+1}{j}\left(x+\frac{k+1}{2}-j\right)_+^k,\quad -\infty<x<\infty \tag{5.26}$$

与式(5.25)相对照可知，$\Omega_k(x)$是定义在区间$(-\infty,\infty)$上以

$$\tilde{x}_j=j-\frac{k+1}{2}\quad (j=0,1,\cdots,k+1)$$

为内节点的k次样条函数，它的内节点个数为$k+2$个，节点的步长为1。

定义 由式(5.26)表示的函数$\Omega_k(x)$称为步长为1、内节点等距的k次B样条。

利用等式$(x-c)(x-c)_+^{k-1}=(x-c)_+^k$可证明$k$次B样条(5.26)具有下列递推关系

$$\begin{aligned}\Omega_k(x)=&\frac{1}{k}\left(x+\frac{k+1}{2}\right)\Omega_{k-1}\left(x+\frac{1}{2}\right)-\\&\frac{1}{k}\left(x-\frac{k+1}{2}\right)\Omega_{k-1}\left(x-\frac{1}{2}\right)\quad (k=2,3,\cdots)\end{aligned} \tag{5.27}$$

利用递推关系(5.27)和数学归纳法又可证明k次B样条(5.26)具有下列性质：

(1) $\Omega_k(-x)=\Omega_k(x)$。

(2) 当$|x|\geqslant\frac{k+1}{2}$时，$\Omega_k(x)\equiv 0$；当$|x|<\frac{k+1}{2}$时，$\Omega_k(x)>0$。

例如，

$$\Omega_1(x)=(x+1)_+-2x_++(x-1)_+=\begin{cases}0, & |x|\geqslant 1\\ 1-|x|, & |x|<1\end{cases}$$

它的内节点为$\tilde{x}_0=-1,\tilde{x}_1=0,\tilde{x}_2=1$，它的图形见图5-1。

$$\Omega_2(x)=\frac{1}{2}\left(x+\frac{3}{2}\right)_+^2-\frac{3}{2}\left(x+\frac{1}{2}\right)_+^2+\frac{3}{2}\left(x-\frac{1}{2}\right)_+^2-\frac{1}{2}\left(x-\frac{3}{2}\right)_+^2=$$

$$\begin{cases}0, & |x|\geqslant\frac{3}{2}\\ -x^2+\frac{3}{4}, & |x|<\frac{1}{2}\\ \frac{1}{2}x^2-\frac{3}{2}|x|+\frac{9}{8}, & \frac{1}{2}\leqslant|x|\leqslant\frac{3}{2}\end{cases}$$

它的内节点为$\tilde{x}_j=j-\frac{3}{2}\ (j=0,1,2,3)$，它的图形见图5-2。

$$\begin{aligned}\Omega_3(x)=&\frac{1}{6}(x+2)_+^3-\frac{2}{3}(x+1)_+^3+x_+^3-\\&\frac{2}{3}(x-1)_+^3+\frac{1}{6}(x-2)_+^3=\\&\begin{cases}0, & |x|\geqslant 2\\ \frac{1}{2}|x|^3-x^2+\frac{2}{3}, & |x|\leqslant 1\\ -\frac{1}{6}|x|^3+x^2-2|x|+\frac{4}{3}, & 1<|x|<2\end{cases}\end{aligned}$$

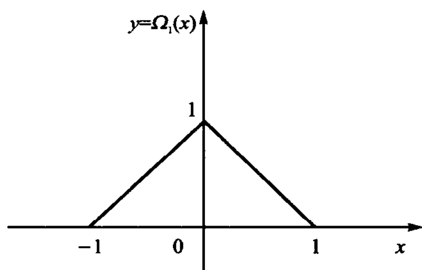


图 5-1　一次 B 样条

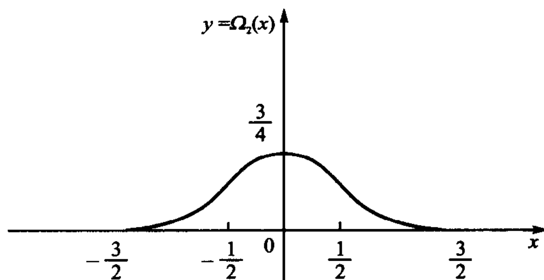


图 5-2　二次 B 样条

它的内节点为 $\tilde{x}_j=j-2\ (j=0,1,2,3,4)$，它的图形见图 5-3。

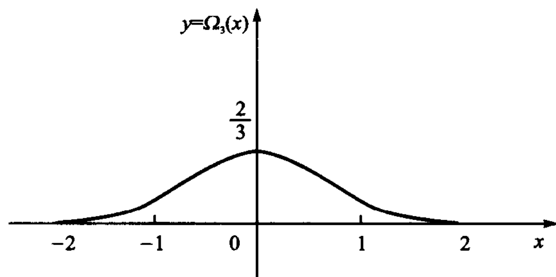


图 5-3　三次 B 样条

设区间$[a,b]$上的样条节点等距，即$[a,b]$上的分划 π 为

$$x_i=a+ih\quad(i=0,1,\cdots,n)\quad h=\frac{b-a}{n}$$

固定某个 i，用$\frac{1}{h}(x-x_{i-\frac{k-1}{2}})$代替 $\Omega_k(x)$中的 x，得到

$$\Omega_k\left(\frac{x-x_{i-\frac{k-1}{2}}}{h}\right)=\frac{1}{k!h^k}\sum_{j=0}^{k+1}(-1)^j\binom{k+1}{j}(x-x_{i-k+j})_+^k,\quad -\infty<x<\infty\qquad(5.28)$$

这是定义在区间$(-\infty,\infty)$上以 $x_{i-k+j}(j=0,1,\cdots,k+1)$为内节点的 k 次样条。称式(5.28)为内节点等距、步长为 h 的 k 次 B 样条。在 Oxy 坐标系中，把 $\Omega_k(x)$的图形按照 $x_{i-\frac{k-1}{2}}$和 h 的数值进行平移和压缩(拉伸)可得到 k 次 B 样条(5.28)的图形，此图形以直线 $x=x_{i-\frac{k-1}{2}}$为对称轴。由 $\Omega_k(x)$的性质可推知

$$\Omega_k\left(\frac{x-x_{i-\frac{k-1}{2}}}{h}\right)\begin{cases}\equiv 0, & x\overline{\in}(x_{i-k},x_{i+1})\\ >0, & x\in(x_{i-k},x_{i+1})\end{cases}$$

考察函数组

$$\mathscr{L}=\left\{\Omega_k\left(\frac{x-x_{j-\frac{k-1}{2}}}{h}\right),a\leqslant x\leqslant b(j=0,1,\cdots,n+k-1)\right\}$$

易知，$\mathscr{L}\subset\mathscr{D}_{k,\pi}$，并且 $\mathscr{L}$ 在区间$[a,b]$上线性无关(证明从略)，$\mathscr{L}$ 所含的函数个数又恰好等于 $\mathscr{D}_{k,\pi}$的维数 $n+k$，所以函数组 $\mathscr{L}$ 是空间 $\mathscr{D}_{k,\pi}$的一组基底。于是有

$$\mathscr{D}_{k,\pi}=\mathrm{Span}\{\mathscr{L}\}$$

定义在$[a,b]$上对应于节点等距分布的分划π的k次样条$s(x)$就可表示为

$$s(x)=\sum_{j=0}^{n+k-1}c_j\Omega_k\left(\frac{x-x_{j-\frac{k-1}{2}}}{h}\right),\quad a\leqslant x\leqslant b \tag{5.29}$$

当$n=4$时，空间$\mathscr{D}_{1,\pi}$的基底图形见图5-4，空间$\mathscr{D}_{3,\pi}$的基底图形见图5-5。

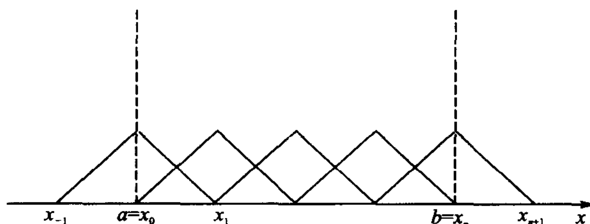


图5-4 空间$\mathscr{D}_{1,\pi}$的基底

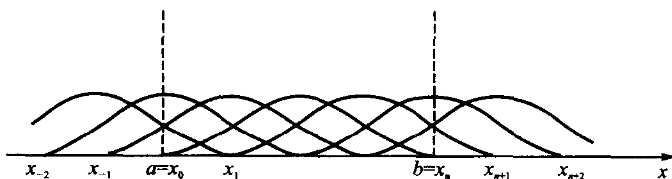


图5-5 空间$\mathscr{D}_{3,\pi}$的基底

5.3.2 三次样条插值问题

由于三次样条具有连续二阶导数，其曲线的光滑性好，所以，在工程技术中通常使用三次样条作为插值函数。

定义 设有区间$[a,b]$上的一个分划

$$\pi:\ a=x_0<x_1<\cdots<x_{n-1}<x_n=b$$

函数$f(x)$在各节点处的值为$y_i=f(x_i)(i=0,1,\cdots,n)$，如果三次样条函数$s(x)\in\mathscr{D}_{3,\pi}$满足条件

$$s(x_i)=y_i\quad(i=0,1,\cdots,n) \tag{5.30}$$

则称$s(x)$为函数$f(x)$的三次样条插值函数。

根据5.3.1小节的讨论，要从空间$\mathscr{D}_{3,\pi}$中确定一个三次样条函数$s(x)$，需要$n+3$个条件。现在式(5.30)提供了$n+1$个条件，还差两个条件。这个事实说明，定义在$[a,b]$上对应于分划π并满足插值条件(5.30)的三次样条插值函数仍然是个函数族。要想从这个函数族中唯一确定一个函数，还要给出两个条件。这两个条件就是在边界节点x_0和x_n处给出导数的约束，称为边界条件。

第一种边界条件：给定两边界节点处的二阶导数$y_0''=f''(x_0)$，$y_n''=f''(x_n)$，并要求$s(x)$满足

$$s''(x_0)=y_0'',\quad s''(x_n)=y_n'' \tag{5.31}$$

特别地，若$y_0''=y_n''=0$，则所得的样条称为自然样条，它表示两端点的约束状态是简支的情形。

第二种边界条件：给定两边界节点处的一阶导数$y_0'=f'(x_0)$，$y_n'=f'(x_n)$，并要求$s(x)$满足

$$s'(x_0) = y'_0, \quad s'(x_n) = y'_n \tag{5.32}$$

第三种边界条件：要求 $s(x)$满足以下的周期性条件

$$s'(x_0^+) = s'(x_n^-), \quad s''(x_0^+) = s''(x_n^-) \tag{5.33}$$

这种边界条件只适合于被插值函数 $f(x)$是以 $x_n - x_0$ 为周期的周期函数。

第一种和第二种边界条件还可以互相搭配产生新的边界条件。

综上所述，三次样条插值问题就是：

给定区间$[a,b]$上的一个分划

$$\pi: \quad a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

$y_i = f(x_i)(i=0,1,\cdots,n)$，又给定一种边界条件。在空间 $\mathscr{D}_{3,\pi}$中求一个三次样条函数$s(x)$，使它满足插值条件(5.30)和给定的边界条件。

定理 5.4　三次样条插值问题的解存在且唯一。

证明从略。

对于在区间$[a,b]$上的连续函数 $f(x)$，记

$$\| f \|_\infty = \max_{a \leqslant x \leqslant b} | f(x) |$$

称$\| f \|_\infty$为函数 $f(x)$的∞-范数。

定理 5.5　设 $f(x)$在区间$[a,b]$上有四阶连续导数，$s(x)$是关于第一或第二种边界条件的三次样条插值问题的解，记 $h_i = x_i - x_{i-1}$，$h = \max\limits_{1 \leqslant i \leqslant n} h_i$，则有估计式

$$\| f^{(m)} - s^{(m)} \|_\infty \leqslant \alpha_m \| f^{(4)} \|_\infty h^{4-m} \quad (m = 0,1,2)$$

其中 $\alpha_0, \alpha_1, \alpha_2$ 都是与 f 和 h 无关的常数。

证明从略。

从定理 5.5 可看出，当 $h \to 0$ 时(因而 $n \to \infty$)，关于第一种和第二种边界条件的三次样条插值函数 $s(x)$及其一阶导数 $s'(x)$、二阶导数 $s''(x)$在区间$[a,b]$上分别一致收敛于 $f(x)$，$f'(x)$和$f''(x)$。

5.3.3　B 样条为基底的三次样条插值函数

设区间$[a,b]$上的分划 π 为

$$x_i = a + ih \quad (i = 0,1,\cdots,n) \quad h = \frac{b-a}{n}$$

则由式(5.29)可知，对应于分划 π 的三次样条插值函数可表示为

$$s(x) = \sum_{j=0}^{n+2} c_j \Omega_3 \left(\frac{x - x_{j-1}}{h} \right), \quad a \leqslant x \leqslant b \tag{5.34}$$

对于第一种边界条件的三次样条插值问题，根据条件(5.30)和式(5.31)，可得

$$\begin{cases} s''(x_0) = \dfrac{1}{h^2} \displaystyle\sum_{j=0}^{n+2} c_j \Omega_3''(1-j) = y_0'' \\ s(x_i) = \displaystyle\sum_{j=0}^{n+2} c_j \Omega_3(i-j+1) = y_i \quad (i = 0,1,\cdots,n) \\ s''(x_n) = \dfrac{1}{h^2} \displaystyle\sum_{j=0}^{n+2} c_j \Omega_3''(n-j+1) = y_n'' \end{cases} \tag{5.35}$$

根据 $\Omega_3(x)$的表达式，可知

$$\Omega_3(0)=\frac{2}{3},\quad \Omega_3(\pm 1)=\frac{1}{6},\quad \Omega_3(x)=0,\quad |x|\geqslant 2$$

$$\Omega_3''(0)=-2,\quad \Omega_3''(\pm 1)=1,\quad \Omega_3''(x)=0,\quad |x|\geqslant 2$$

于是,方程组(5.35)的具体形式是

$$\begin{cases} c_0-2c_1+c_2=h^2y_0'' \\ \dfrac{1}{6}c_i+\dfrac{2}{3}c_{i+1}+\dfrac{1}{6}c_{i+2}=y_i \quad (i=0,1,\cdots,n) \\ c_n-2c_{n+1}+c_{n+2}=h^2y''_n \end{cases} \tag{5.36}$$

方程组(5.36)是一个关于 $c_0,c_1,\cdots,c_{n+2}$ 的 $n+3$ 元线性方程组。由方程组(5.36)的第一和第二个方程解出

$$\begin{cases} c_0=2c_1-c_2+h^2y_0'' \\ c_1=y_0-\dfrac{h^2}{6}y_0'' \end{cases} \tag{5.37}$$

由第 $n+2$ 和第 $n+3$ 个方程解出

$$\begin{cases} c_{n+1}=y_n-\dfrac{h^2}{6}y_n'' \\ c_{n+2}=2c_{n+1}-c_n+h^2y_n'' \end{cases} \tag{3.38}$$

又在方程组(5.36)中消去 c_0,c_1,c_{n+1},c_{n+2},得 $n-1$ 元线性方程组

$$\boldsymbol{Ac}=\boldsymbol{d} \tag{5.39}$$

其中

$$\boldsymbol{A}=\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix},\quad \boldsymbol{d}=\begin{bmatrix} 6y_1-y_0+\dfrac{h^2}{6}y_0'' \\ 6y_2 \\ 6y_3 \\ \vdots \\ 6y_{n-2} \\ 6y_{n-1}-y_n+\dfrac{h^2}{6}y_n'' \end{bmatrix}$$

$$\boldsymbol{c}=(c_2,c_3,\cdots,c_n)^{\mathrm{T}}$$

首先从方程(5.39)解出 $c_2,c_3,\cdots,c_n$,由于 **A** 是主对角线按行严格占优阵,故方程组(5.39)有唯一解。再从式(5.37)、式(5.38)算出 c_0,c_1,c_{n-1},c_{n+2}。把所求出的系数 $c_j(j=0,1,\cdots,n+2)$ 代入式(5.34)就得到关于第一种边界条件的三次样条插值函数 $s(x)$。

对于第二种边界条件的三次样条插值问题,则由条件(5.30)和式(5.32)可推出关系式

$$\begin{cases} c_0=c_2-2hy_0' \\ c_{n+2}=c_n+2hy_n' \end{cases} \tag{5.40}$$

以及关于 $c_1,c_2,\cdots,c_{n+1}$ 的 $n+1$ 元线性方程组

$$\boldsymbol{Bc}=\boldsymbol{v} \tag{5.41}$$

其中

$$
\boldsymbol{B} = \begin{bmatrix} 4 & 2 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 2 & 4 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} 6y_0 + 2hy'_0 \\ 6y_1 \\ 6y_2 \\ \vdots \\ 6y_{n-1} \\ 6y_n - 2hy'_n \end{bmatrix}
$$

$$
\boldsymbol{c} = (c_1, c_2, \cdots, c_{n+1})^{\mathrm{T}}
$$

由方程组(5.41)解出 $c_1, c_2, \cdots, c_{n+1}$，再由式(5.40)算出 c_0 和 c_{n+2}，把所得的系数代入式(5.34)就得到关于第二种边界条件的三次样条插值函数 $s(x)$。

对于第三种边界条件的三次样条插值问题，由式(5.30)和式(5.33)可推出关系式

$$
c_{n+2} = c_2, \quad c_1 = c_{n+1}, \quad c_0 = c_n \tag{5.42}
$$

以及 n 元线性方程组

$$
\boldsymbol{Pc} = \boldsymbol{u} \tag{5.43}
$$

其中

$$
\boldsymbol{P} = \begin{bmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} 6y_1 \\ 6y_2 \\ \vdots \\ 6y_{n-1} \\ 6y_0 \end{bmatrix}
$$

$$
\boldsymbol{c} = (c_2, c_3, \cdots, c_{n+1})^{\mathrm{T}}
$$

从方程组(5.43)解出 $c_2, c_3, \cdots, c_{n+1}$，由式(5.42)直接确定 c_{n+2}, c_1, c_0，把所得系数代入式(5.34)就得到关于第三种边界条件的三次样条插值函数 $s(x)$。

例 4 给定数表

x	1	2	3
$f(x)$	2	4	8
$f'(x)$	1.386 3		5.545 2

求以 $x_0=1, x_1=2, x_2=3$ 为节点的三次样条函数 $s(x)$，使其满足条件

$$
s(x_i) = f(x_i) \quad (i = 0,1,2)
$$

$$
s'(x_0) = f'(x_0), \quad s'(x_2) = f'(x_2)
$$

解 这是属于第二种边界条件的三次样条插值问题，且是 $n=2$ 的情形。首先求解方程组(5.41)。这时 $h=1$，并且

$$
\boldsymbol{v} = \begin{bmatrix} 6y_0 + 2hy'_0 \\ 6y_1 \\ 6y_2 - 2hy'_2 \end{bmatrix} = \begin{bmatrix} 14.772\,6 \\ 24 \\ 36.909\,6 \end{bmatrix}
$$

方程组(5.41)成为

$$
\begin{bmatrix} 4 & 2 & 0 \\ 1 & 4 & 1 \\ 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 14.772\,6 \\ 24 \\ 36.909\,6 \end{bmatrix}
$$

解得

$$c_1 = 1.846\ 57, \quad c_2 = 3.693\ 17, \quad c_3 = 7.380\ 81$$

又由式(5.40)，得

$$c_0 = c_2 - 2hy'_0 = 0.920\ 570$$

$$c_4 = c_2 + 2hy'_2 = 14.783\ 6$$

于是，由式(5.34)得到所求的三次样条插值函数 $s(x)$ 为

$$\begin{aligned} s(x) = & 0.920\ 570\Omega_3(x) + 1.846\ 57\Omega_3(x-1) + \\ & 3.693\ 17\Omega_3(x-2) + 7.380\ 81\Omega_3(x-3) + \\ & 14.783\ 6\Omega_3(x-4), \quad 1 \leqslant x \leqslant 3 \end{aligned}$$

根据三次 B 样条 $\Omega_3(x)$ 的表达式，$s(x)$ 又可整理为

$$\begin{aligned} s(x) = & 0.920\ 594 + 0.925\ 915x + 0.000\ 077\ 001x^2 + \\ & 0.153\ 406x^3 + 0.158\ 945(x-2)_+^3, \quad 1 \leqslant x \leqslant 3 \end{aligned}$$

5.3.4 三弯矩法求三次样条插值函数

设区间 $[a,b]$ 上的分划

$$\pi: \quad a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

是任意的。对于这种情形，三次样条插值函数 $s(x)$ 可以采用式(5.25)的形式，即

$$s(x) = \sum_{j=1}^{3} a_j x^j + \frac{1}{3!}\sum_{j=1}^{n-1} c_j (x-x_j)_+^3, \quad a \leqslant x \leqslant b \tag{5.44}$$

然后根据插值条件(5.30)和某种边界条件确定式(5.44)中的 $n+3$ 个系数 $a_j(j=0,1,2,3)$，$c_j(j=1,2,\cdots,n-1)$。但此法需要求解一个系数矩阵较为复杂的 $n+3$ 元线性方程组，一般不用此法。还有多种其他方法可选择，这里介绍其中的一种，被称为三弯矩法。

记

$$M_i = s''(x_i) \quad (i = 0,1,\cdots,n)$$

由于 $s(x)$ 在子区间 $[x_{i-1},x_i]$ 上的表达式是次数不高于 3 的代数多项式，所以 $s''(x)$ 在该子区间上是线性函数，并且有

$$s''(x) = \frac{x-x_i}{-h_i}M_{i-1} + \frac{x-x_{i-1}}{h_i}M_i \tag{5.45}$$

其中 $h_i = x_i - x_{i-1}$。将等式(5.45)积分两次，得到

$$s(x) = \frac{M_{i-1}}{6h_i}(x_i - x)^3 + \frac{M_i}{6h_i}(x-x_{i-1})^3 + c_1 x + c_2 \tag{5.46}$$

利用插值条件

$$s(x_{i-1}) = y_{i-1}, \quad s(x_i) = y_i$$

定出积分常数 c_1 和 c_2，然后代入式(5.46)中并整理，得到

$$\begin{aligned} s(x) = & \frac{M_{i-1}}{6h_i}(x_i - x)^3 + \frac{M_i}{6h_i}(x-x_{i-1})^3 + \\ & \left(\frac{y_{i-1}}{h_i} - \frac{M_{i-1}}{6}h_i\right)(x_i - x) + \\ & \left(\frac{y_i}{h_i} - \frac{M_i}{6}h_i\right)(x-x_{i-1}), \quad x_{i-1} \leqslant x \leqslant x_i \quad (i = 1,2,\cdots,n) \end{aligned} \tag{5.47}$$

只要能把 $M_i(i=0,1,\cdots,n)$ 求出来，所求的三次样条插值函数 $s(x)$ 在各个子区间上的表达式

就由(5.47)确定。为了求 M_i,需要利用 $s'(x)$ 在各个内节点 $x_i(i=1,2,\cdots,n-1)$ 处连续的条件。由式(5.47)得

$$s'(x)=-\frac{M_{i-1}}{2h_i}(x_i-x)^2+\frac{M_i}{2h_i}(x-x_{i-1})^2+$$

$$\frac{y_i-y_{i-1}}{h_i}-\frac{h_i}{6}(M_i-M_{i-1}),\quad x_{i-1}<x<x_i\quad(i=1,2,\cdots,n)\tag{5.48}$$

由式(5.48)可知

$$s'(x_i^-)=\frac{h_i}{6}M_{i-1}+\frac{h_i}{3}M_i+\frac{y_i-y_{i-1}}{h_i}$$

$$s'(x_i^+)=-\frac{h_{i+1}}{3}M_i-\frac{h_{i+1}}{6}M_{i+1}+\frac{y_{i+1}-y_i}{h_{i+1}}$$

因 $s'(x)$ 在 x_i 处连续,故应有

$$s'(x_i^-)=s'(x_i^+)\quad(i=1,2,\cdots,n-1)$$

由此得到 $n-1$ 个方程(称为三弯矩方程)

$$\gamma_iM_{i-1}+2M_i+\alpha_iM_{i+1}=\beta_i\quad(i=1,2,\cdots,n-1)\tag{5.49}$$

其中

$$\alpha_i=\frac{h_{i+1}}{h_i+h_{i+1}},\quad\gamma_i=1-\alpha_i$$

$$\beta_i=\frac{6}{h_i+h_{i+1}}\left(\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{y_i-y_{i-1}}{h_i}\right)$$

式(5.49)是关于未知数 $M_0,M_1,\cdots,M_n$ 的线性方程组,要唯一确定这 $n+1$ 个未知数须增加两个方程。这就需要边界条件。

对于第一种边界条件,相当于增加以下两个方程

$$2M_0+\alpha_0M_1=\beta_0\tag{5.50}$$

$$\gamma_nM_{n-1}+2M_n=\beta_n\tag{5.51}$$

其中

$$\alpha_0=0,\quad\beta_0=2y_0'',\quad\gamma_n=0,\quad\beta_n=2y_n''$$

将方程(5.50),(5.49),(5.51)合在一起,构成了关于 $M_0,M_1,\cdots,M_n$ 的 $n+1$ 元线性方程组

$$\begin{bmatrix}2&\alpha_0&&&\\\gamma_1&2&\alpha_1&&\\&\ddots&\ddots&\ddots&\\&&\gamma_{n-1}&2&\alpha_{n-1}\\&&&\gamma_n&2\end{bmatrix}\begin{bmatrix}M_0\\M_1\\\vdots\\M_{n-1}\\M_n\end{bmatrix}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_{n-1}\\\beta_n\end{bmatrix}\tag{5.52}$$

它的系数矩阵是主对角线按行严格占优阵,故它有唯一解。

对于第二种边界条件,即

$$s'(x_0)=y_0',\quad s'(x_n)=y_n'$$

由式(5.48)可得如下两个方程

$$-\frac{h_1}{3}M_0-\frac{h_1}{6}M_1+\frac{y_1-y_0}{h_1}=y_0'$$

$$\frac{h_n}{6}M_{n-1}+\frac{h_n}{3}M_n+\frac{y_n-y_{n-1}}{h_n}=y_n'$$

经过整理，这两个方程也分别写成方程(5.50)和方程(5.51)的形式，但这里

$$\alpha_0 = 1, \quad \beta_0 = \frac{6}{h_1}\left(\frac{y_1 - y_0}{h_1} - y'_0\right)$$

$$\gamma_n = 1, \quad \beta_n = \frac{6}{h_n}\left(y'_n - \frac{y_n - y_{n-1}}{h_n}\right)$$

所得的关于 $M_0, M_1, \cdots, M_n$ 的线性方程组仍然是方程组(5.52)的形式。

对于第三种边界条件，则由

$$s'(x_0^+) = s'(x_n^-), \quad s''(x_0^+) = s''(x_n^-)$$

得到两个方程

$$-\frac{h_1}{3}M_0 - \frac{h_1}{6}M_1 + \frac{y_1 - y_0}{h_1} = \frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n + \frac{y_n - y_{n-1}}{h_n}$$

$$M_0 = M_n$$

从这两个方程中消去 M_0，并经整理，化简为

$$\alpha_n M_1 + \gamma_n M_{n-1} + 2M_n = \beta_n \tag{5.53}$$

其中

$$\alpha_n = \frac{h_1}{h_1 + h_n}, \quad \gamma_n = 1 - \alpha_n$$

$$\beta_n = \frac{6}{h_1 + h_n}\left(\frac{y_1 - y_0}{h_1} - \frac{y_n - y_{n-1}}{h_n}\right)$$

由方程(5.49)及式(5.53)构成以 $M_1, M_2, \cdots, M_n$ 为未知数的线性方程组

$$\begin{bmatrix} 2 & \alpha_1 & & & \gamma_1 \\ \gamma_2 & 2 & \alpha_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-1} & 2 & \alpha_{n-1} \\ \alpha_n & & & \gamma_n & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-1} \\ \beta_n \end{bmatrix} \tag{5.54}$$

方程组(5.54)的系数矩阵也是非奇异的，因而此方程组也有唯一解。

综上所述，为确定三次样条插值函数 $s(x)$，其计算步骤为：

(1) 根据给定的$(x_i, y_i)(i=0,1,\cdots,n)$以及边界条件，计算关于 $M_0, M_1, \cdots, M_n$ 的线性方程组中的有关参数(系数矩阵的元素和右端项)。

(2) 求解上述线性方程组(可用追赶法)。

(3) 把求出的 $M_0, M_1, \cdots, M_n$ 代入式(5.47)，所得的 $s(x)$就是所求的三次样条插值函数。

把 $M_0, M_1, \cdots, M_n$ 代入式(5.48)还可得到三次样条插值函数的导数 $s'(x)$。在实际应用中，不仅常用三次样条函数 $s(x)$计算被插值函数 $f(x)$的近似值，而且常用 $s'(x)$近似计算 $f'(x)$。

在力学中，二阶导数决定梁的弯矩。由于方程组(5.52)或方程组(5.54)中的每一个方程至多出现相邻三个节点处的 M_i，故称上述方法为三弯矩法。

例 5 用三弯矩法求解例 4 提出的三次样条插值问题。

解 首先计算方程组(5.52)中的有关参数，得到

$$\alpha_0 = 1, \quad \beta_0 = 3.682\,2, \quad \gamma_2 = 1, \quad \beta_2 = 9.271\,2$$

$$\alpha_1 = 0.5, \quad \beta_1 = 6, \quad \gamma_1 = 0.5$$

然后求解方程组

$$\begin{bmatrix} 2 & 1 & 0 \\ 0.5 & 2 & 0.5 \\ 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} M_0 \\ M_1 \\ M_2 \end{bmatrix} = \begin{bmatrix} 3.6822 \\ 6 \\ 9.2712 \end{bmatrix}$$

解得 $M_0=0.92055, M_1=1.8411, M_2=3.71505$，代入式(5.47)，经整理得

$$s(x)=\begin{cases} 0.92055+0.926025x+0.153425x^3, & 1\leqslant x\leqslant 2 \\ -0.35065+2.832825x-0.9534x^2+0.312325x^3, & 2<x\leqslant 3 \end{cases}$$

本例计算过程中没有任何舍入误差，而在例 4 的计算过程中有舍入误差，所以两个结果略有不同。

5.4　三角插值与快速 Fourier 变换

5.4.1　周期函数的三角插值

现在讨论周期函数的三角插值问题，它相当于离散情形的频谱分析。

设 $f(x)\in C(-\infty,\infty)$，并且是以 2π 为周期的实值函数。在区间 $[0,2\pi]$ 上给定 N 个等距节点

$$x_l=\frac{2\pi l}{N}\quad (l=0,1,\cdots,N-1)$$

并已知节点处的函数值 $f(x_l)(l=0,1,\cdots,N-1)$。取三角函数组

$$\mathscr{L}=\{1,\cos x,\sin x,\cdots,\cos nx,\sin nx\}$$

作为插值基函数，其中 $2n+1=N$。在三角多项式集合

$$\mathscr{D}_n=\mathrm{Span}\{\mathscr{L}\}$$

中寻求三角多项式

$$s_n(x)=\frac{a_0}{2}+\sum_{j=1}^{n}(a_j\cos jx+b_j\sin jx) \tag{5.55}$$

使其满足插值条件

$$s_n(x_l)=f(x_l)\quad (l=0,1,\cdots,N-1) \tag{5.56}$$

此问题称为周期函数的三角插值问题。满足条件(5.56)的三角多项式 $s_n(x)$ 称为函数 $f(x)$ 的三角插值多项式或三角插值函数。

易知，三角函数组 $\mathscr{L}$ 在点集 $\left\{x_l=\frac{2\pi l}{N}(l=0,1,\cdots,N-1)\right\}$ 上正交，即向量组

$$\begin{aligned} \boldsymbol{\phi}_1 &= (1,1,\cdots,1)^{\mathrm{T}} \\ \boldsymbol{\phi}_{2k} &= (\cos kx_0,\cos kx_1,\cdots,\cos kx_{N-1})^{\mathrm{T}} \\ \boldsymbol{\phi}_{2k+1} &= (\sin kx_0,\sin kx_1,\cdots,\sin kx_{N-1})^{\mathrm{T}} \end{aligned}$$

$$(k=1,2,\cdots,n)$$

是正交向量组，其原因是下列等式成立

$$\begin{cases} \sum_{l=0}^{N-1} \cos k\,\frac{2\pi l}{N} = 0, \quad \sum_{l=0}^{N-1} \sin k\,\frac{2\pi l}{N} = 0 \quad (k = 1,2,\cdots,n) \\ \sum_{l=0}^{N-1} \cos k\,\frac{2\pi l}{N} \cos j\,\frac{2\pi l}{N} = \begin{cases} 0, & k \neq j \\ \frac{N}{2}, & k = j \end{cases} \quad (k,j = 1,2,\cdots,n) \\ \sum_{l=0}^{N-1} \cos k\,\frac{2\pi l}{N} \sin j\,\frac{2\pi l}{N} = 0 \quad (k,j = 1,2,\cdots,n) \\ \sum_{l=0}^{N-1} \sin k\,\frac{2\pi l}{N} \sin j\,\frac{2\pi l}{N} = \begin{cases} 0, & k \neq j \\ \frac{N}{2}, & k = j \end{cases} \quad (k,j = 1,2,\cdots,n) \end{cases} \tag{5.57}$$

由此可知，函数 $f(x)$ 的三角插值多项式 $s_n(x)$ 存在且唯一。

由条件(5.56)得到关于系数 $a_0, a_1, b_1, \cdots, a_n, b_n$ 的线性方程组

$$\frac{a_0}{2} + \sum_{j=1}^{n}\left(a_j \cos j\,\frac{2\pi l}{N} + b_j \sin j\,\frac{2\pi l}{N}\right) = f\left(\frac{2\pi l}{N}\right) \quad (l = 0,1,\cdots,N-1) \tag{5.58}$$

利用式(5.57)中的等式即可从方程组(5.58)中解出系数 a_j 和 b_j，结果如下：

$$\begin{cases} a_j = \frac{2}{N}\sum_{l=0}^{N-1} f\left(\frac{2\pi l}{N}\right)\cos j\,\frac{2\pi l}{N} \quad (j = 0,1,\cdots,n) \\ b_j = \frac{2}{N}\sum_{l=0}^{N-1} f\left(\frac{2\pi l}{N}\right)\sin j\,\frac{2\pi l}{N} \quad (j = 1,2,\cdots,n) \end{cases} \tag{5.59}$$

于是，$f(x)$ 的三角插值多项式(5.55)被确定。

一般情形下，$f(x)$ 是以 2π 为周期的复值函数，它在 N 个节点 $x_l = \frac{2\pi l}{N}(l=0,1,\cdots,N-1)$ 处的函数值 $f_l = f\left(\frac{2\pi l}{N}\right)(l=0,1,\cdots,N-1)$ 为已知。取复值函数组

$$\varphi_k(x) = \mathrm{e}^{\mathrm{i}kx} = \cos kx + \mathrm{i}\sin kx \quad (k = 0,1,\cdots,N-1)$$

作为插值基函数，其中 $\mathrm{i}=\sqrt{-1}$。此时，复值函数 $f(x)$ 的三角插值多项式 $s_n(x)$ 的形式为

$$s_n(x) = \sum_{k=0}^{N-1} c_k \mathrm{e}^{\mathrm{i}kx} \tag{5.60}$$

由于复向量组

$$\boldsymbol{\Phi}_k = (\varphi_k(x_0), \varphi_k(x_1), \cdots, \varphi_k(x_{N-1}))^{\mathrm{T}} \quad (k = 0,1,\cdots,N-1)$$

满足

$$(\boldsymbol{\Phi}_k, \boldsymbol{\Phi}_j) = \sum_{l=0}^{N-1} \varphi_k(x_l)\,\overline{\varphi_j(x_l)} = \sum_{l=0}^{N-1} \mathrm{e}^{\mathrm{i}(k-j)\frac{2\pi l}{N}} = \begin{cases} 0, & k \neq j \\ N, & k = j \end{cases} \quad (k,j = 0,1,\cdots,N-1) \tag{5.61}$$

所以，复函数组 $\{\varphi_k(x) = \mathrm{e}^{\mathrm{i}kx}(k=0,1,\cdots,N-1)\}$ 在点集 $\left\{x_l = \frac{2\pi l}{N}(l=0,1,\cdots,N-1)\right\}$ 上正交。由此可知，满足插值条件

$$s_n(x_l) = f_l \quad (l = 0,1,\cdots,N-1) \tag{5.62}$$

的三角插值多项式 $s_n(x)$ 存在且唯一。

由条件(5.62)和 $s_n(x)$ 的表达式(5.60)得到关于系数 $c_0, c_1, \cdots, c_{N-1}$ 的线性方程组

$$\sum_{k=0}^{N-1} c_k e^{ik\frac{2\pi l}{N}} = f_l \quad (l = 0,1,\cdots,N-1) \tag{5.63}$$

为求 c_j,用 $e^{-ij\frac{2\pi l}{N}}$ 乘方程组(5.63)第 l 个方程的两端($l=0,1,\cdots,N-1$),再把 N 个方程相加,得

$$\sum_{l=0}^{N-1} \sum_{k=0}^{N-1} c_k e^{i(k-j)\frac{2\pi l}{N}} = \sum_{l=0}^{N-1} f_l e^{-ij\frac{2\pi l}{N}} \tag{5.64}$$

由式(5.64)并根据式(5.61)即可得到

$$c_j = \frac{1}{N} \sum_{l=0}^{N-1} f_l e^{-ij\frac{2\pi l}{N}} \quad (j = 0,1,\cdots,N-1) \tag{5.65}$$

至此,以 2π 为周期的复值函数 $f(x)$ 的三角插值多项式(5.60)已被确定。

把式(5.65)和式(5.63)重新表达如下:

$$c_k = \frac{1}{N} \sum_{l=0}^{N-1} f_l e^{-ikl\frac{2\pi}{N}} \quad (k = 0,1,\cdots,N-1) \tag{5.65$_1$}$$

$$f_l = \sum_{k=0}^{N-1} c_k e^{ikl\frac{2\pi}{N}} \quad (l = 0,1,\cdots,N-1) \tag{5.63$_1$}$$

由 $\{f_l(l=0,1,\cdots,N-1)\}$ 通过公式(5.65)$_1$ 求 $\{c_k(k=0,1,\cdots,N-1)\}$,称为对 $f(x)$ 的离散 Fourier(傅里叶)变换,简称 DFT;反过来,由 $\{c_k(k=0,1,\cdots,N-1)\}$ 通过公式(5.63)$_1$ 求 $\{f_l(l=0,1,\cdots,N-1)\}$,称为离散 Fourier 逆变换。称 $\{c_k\}$ 是 $\{f_l\}$ 的离散频谱。

5.4.2 快速 Fourier 变换

如果直接用公式(5.65)$_1$ 计算 c_k,那么计算全部 $\{c_k\}$ 共需 N^2 次复数乘法运算和 $N(N-1)$ 次复数加法运算。当实际频谱分析中 N 较大时,这个计算量太大了。因此,在相当长的时间内,各种领域的频谱分析问题中,数值方法没有得到广泛应用。直到 20 世纪 60 年代中期提出了快速 Fourier 变换(Fast Fourier Transform,简称 FFT 算法)才使问题得到解决。这种算法的思想是利用函数 $e^{-i\frac{2\pi}{N}}$ 的周期性。

取 $N=2^m$,m 是正整数。记 $W=e^{-i\frac{2\pi}{N}}$,对任何整数 r 均有 $W^{rN}=1$。公式(5.65)$_1$ 可写成

$$c_k = \sum_{l=0}^{N-1} \frac{1}{N} f_l W^{kl} \quad (k = 0,1,\cdots,N-1) \tag{5.66}$$

以下以 $N=8(m=3)$ 为例说明 FFT 的思想。

用二进制数表示 k 和 l,即

$$k = 2^2 k_2 + 2^1 k_1 + 2^0 k_0 = (k_2, k_1, k_0)$$
$$l = 2^2 l_2 + 2^1 l_1 + 2^0 l_0 = (l_2, l_1, l_0)$$

其中 $k_2, k_1, k_0, l_2, l_1, l_0$ 只取 0 和 1 两个值。又记

$$c(k_2, k_1, k_0) = c_k, \quad a_0(l_2, l_1, l_0) = \frac{1}{N} f_l$$

例如,$c_2 = c(0,1,0)$,$\frac{1}{N} f_4 = a_0(1,0,0)$。于是,式(5.66)可表示为

$$c(k_2, k_1, k_0) = \sum_{l=0}^{7} \frac{1}{N} f_l W^{kl} =$$

$$\sum_{l_0=0}^{1}\sum_{l_1=0}^{1}\sum_{l_2=0}^{1}a_0(l_2,l_1,l_0)W^{(k_2,k_1,k_0)(l_2,l_1,l_0)} \quad (k_0,k_1,k_2=0,1) \tag{5.67}$$

由于

$$(k_2,k_1,k_0)(l_2,l_1,l_0)=2^4k_2l_2+2^3(k_2l_1+k_1l_2)+ \\ 2^2(k_2l_0+k_1l_1+k_0l_2)+2^1(k_1l_0+k_0l_1)+2^0k_0l_0$$

并注意到 $W^0=W^8=W^{16}=1$，所以

$$W^{(k_2,k_1,k_0)(l_2,l_1,l_0)}=W^{k_0(l_2,l_1,l_0)}W^{k_1(l_1,l_0,0)}W^{k_2(l_0,0,0)} \tag{5.68}$$

把式(5.68)代入式(5.67)，得

$$c(k_2,k_1,k_0)=\sum_{l_0=0}^{1}\left\{\sum_{l_1=0}^{1}\left[\sum_{l_2=0}^{1}a_0(l_2,l_1,l_0)W^{k_0(l_2,l_1,l_0)}\right]W^{k_1(l_1,l_0,0)}\right\}\cdot \\ W^{k_2(l_0,0,0)} \quad (k_0,k_1,k_2=0,1) \tag{5.69}$$

把式(5.69)分解成下列的递推形式：

$$\begin{cases} a_0(l_2,l_1,l_0)=\dfrac{1}{N}f_l \\ a_1(l_1,l_0,k_0)=\displaystyle\sum_{l_2=0}^{1}a_0(l_2,l_1,l_0)W^{k_0(l_2,l_1,l_0)} \\ a_2(l_0,k_1,k_0)=\displaystyle\sum_{l_1=0}^{1}a_1(l_1,l_0,k_0)W^{k_1(l_1,l_0,0)} \\ a_3(k_2,k_1,k_0)=\displaystyle\sum_{l_0=0}^{1}a_2(l_0,k_1,k_0)W^{k_2(l_0,0,0)} \\ c_k=c(k_2,k_1,k_0)=a_3(k_2,k_1,k_0) \\ \qquad \text{其中 } l_p=0,1;k_p=0,1 \quad (p=0,1,2) \end{cases} \tag{5.70}$$

递推公式(5.70)就是 $N=8$ 时的 FFT 算法。从算法中看出，数集$\{a_1(l_1,l_0,k_0)\}$共有 2^3 个数，求出这 2^3 个数须要做 2^3 次复数乘法运算和 2^3 次复数加法运算。其余两个数集$\{a_2(l_0,k_1,k_0)\}$和$\{a_3(k_2,k_1,k_0)\}$也是如此。于是，为求出数集$\{c_k(k=0,1,\cdots,7)\}$，共须做复数乘法和加法运算的次数都是 $3\times 2^3=3N$。

对于 $N=2^m$ 时的 FFT 算法如下：

$$\begin{cases} a_0(l_{m-1},\cdots,l_1,l_0)=\dfrac{1}{N}f_l \\ a_j(l_{m-j-1},\cdots,l_1,l_0,k_{j-1},\cdots,k_1,k_0)= \\ \qquad \displaystyle\sum_{l_{m-j}=0}^{1}a_{j-1}(l_{m-j},\cdots,l_1,l_0,k_{j-2},\cdots,k_1,k_0)\cdot \\ \qquad W^{k_{j-1}(l_{m-j},\cdots,l_1,l_0,0,\cdots,0)} \quad (j=1,2,\cdots,m) \\ c_k=c(k_{m-1},\cdots,k_1,k_0)=a_m(k_{m-1},\cdots,k_1,k_0) \\ \qquad \text{其中 } l_p=0,1;k_p=0,1 \quad (p=0,1,\cdots,m-1) \end{cases} \tag{5.71}$$

算法(5.71)称为以 2 为底的 FFT 算法，这个算法从数集$\left\{\dfrac{1}{N}f_l(l=0,1,\cdots,N-1)\right\}$算出数集$\{c_k(k=0,1,\cdots,N-1)\}$共须做 mN 次复数乘法运算和复数加法运算。这个计算量比直接使

用公式(5.65)$_1$ 计算所需的 N^2 次运算次数是很大的节省。N 越大，FFT 算法(5.71)的相对效益越高。此外，还可以对算法(5.71)做些改进，进一步减少计算量，对此，本书就不再讨论了。

5.5 正交多项式

5.5.1 正交多项式概念与性质

定义 若在区间(a,b)(有限或无限)上非负的函数 $\rho(x)$满足

(1) 对一切整数 $n\geqslant 0$，$\int_a^b x^n\rho(x)\mathrm{d}x$ 存在；

(2) 对区间(a,b)上非负连续函数 $f(x)$，若 $\int_a^b \rho(x)f(x)\mathrm{d}x=0$，则在$(a,b)$上 $f(x)\equiv 0$，

那么，就称 $\rho(x)$为区间(a,b)上的权函数。

常见的权函数有

$$\begin{aligned}
&\rho(x)\equiv 1, && a\leqslant x\leqslant b\\
&\rho(x)=\frac{1}{\sqrt{1-x^2}}, && -1<x<1\\
&\rho(x)=\sqrt{1-x^2}, && -1\leqslant x\leqslant 1\\
&\rho(x)=\mathrm{e}^{-x}, && 0\leqslant x<\infty\\
&\rho(x)=\mathrm{e}^{-x^2}, && -\infty<x<\infty
\end{aligned}$$

定义 给定 $f(x),g(x)\in C[a,b]$，$\rho(x)$是(a,b)上的权函数，称

$$(f,g)=\int_a^b \rho(x)f(x)g(x)\mathrm{d}x$$

为函数 $f(x)$与 $g(x)$在$[a,b]$上的内积。

内积有以下一些简单性质：

(1) $(f,g)=(g,f)$；

(2) $(kf,g)=(f,kg)=k(f,g)$，k 为常数；

(3) $(f_1+f_2,g)=(f_1,g)+(f_2,g)$；

(4) 若在$[a,b]$上 $f(x)\not\equiv 0$，则$(f,f)>0$。

这些性质都可由定积分的性质推出。

定义 若内积

$$(f,g)=\int_a^b \rho(x)f(x)g(x)\mathrm{d}x=0$$

则称 $f(x)$与 $g(x)$在区间$[a,b]$上带权 $\rho(x)$正交。若函数系$\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x),\cdots\}$满足

$$(\varphi_i,\varphi_j)=\int_a^b \rho(x)\varphi_i(x)\varphi_j(x)\mathrm{d}x=\begin{cases}0, & i\neq j\\ a_i>0, & i=j\end{cases}$$

则称$\{\varphi_k(x)\}$是$[a,b]$上带权 $\rho(x)$的正交函数系。特别地，若 $\varphi_k(x)(k=0,1,\cdots)$是最高次项系数不为零的 k 次多项式，则称$\{\varphi_k(x)\}$是$[a,b]$上带权 $\rho(x)$的正交多项式系。

容易证明，在$[a,b]$上带权$\rho(x)$的正交函数系一定是在$[a,b]$上线性无关的函数系，而不论$\rho(x)$是什么。

定理 5.6　设$\varphi_k(x)(k=0,1,\cdots)$是最高次项系数不为零的$k$次多项式，则多项式系$\{\varphi_k(x)\}$为$[a,b]$上带权$\rho(x)$的正交多项式系的充分必要条件是对任何次数不高于$k-1$的多项式$q(x)$，总有

$$\int_a^b \rho(x)q(x)\varphi_k(x)\mathrm{d}x = 0 \quad (k = 1,2,\cdots) \tag{5.72}$$

证

必要性　任何次数不高于$k-1$的多项式$q(x)(k\geqslant 1)$总可表示为某一组0次，1次，…，$k-1$次多项式的线性组合，特别地，可表示为$\varphi_0(x),\varphi_1(x),\cdots,\varphi_{k-1}(x)$的线性组合

$$q(x) = \sum_{j=0}^{k-1} c_j\varphi_j(x)$$

因而有

$$\int_a^b \rho(x)q(x)\varphi_k(x)\mathrm{d}x = \sum_{j=0}^{k-1} c_j\int_a^b \rho(x)\varphi_j(x)\varphi_k(x)\mathrm{d}x$$

因$j\neq k$，故上式右端每个积分皆等于零，所以式(5.72)成立。

充分性　因对任何次数不高于$k-1$的多项式$q(x)$，式(5.72)成立，所以，对于$\varphi_j(x)$ $(j=0,1,\cdots,k-1)$应有

$$\int_a^b \rho(x)\varphi_j(x)\varphi_k(x)\mathrm{d}x = 0 \quad (k = 1,2,\cdots)$$

即

$$(\varphi_j,\varphi_k) = 0,\quad j \neq k$$

又因$\varphi_k(x)(k=0,1,\cdots)$是最高次项系数不为零的$k$次多项式，故$\varphi_k(x)\not\equiv 0(x\in[a,b])$，因而有

$$(\varphi_k,\varphi_k) > 0 \quad (k = 0,1,\cdots)$$

根据定义，$\{\varphi_k(x)\}$是$[a,b]$上带权$\rho(x)$的正交多项式系。

证毕。

正交多项式还有以下的性质。

性质 1　设$\{\varphi_k(x)\}$是$[a,b]$上带权$\rho(x)$的正交多项式系，则$\{c_k\varphi_k(x)\}$也是$[a,b]$上带权$\rho(x)$的正交多项式系，其中$c_k(k=0,1,\cdots)$是非零常数。

此性质是明显的。

性质 2　区间$[a,b]$上带权$\rho(x)$的正交多项式系，在各个多项式的最高次项系数为1的情形下是唯一的。

证　设$\{g_k(x)\}$和$\{\varphi_k(x)\}$都是$[a,b]$上带权$\rho(x)$的正交多项式系，并且对任何$k\geqslant 0$，$g_k(x)$和$\varphi_k(x)$的x^k项系数为1。

当$k=0$时，$g_0(x)\equiv 1,\varphi_0(x)\equiv 1$，所以$g_0(x)\equiv\varphi_0(x)$。

当$k\geqslant 1$时，$g_k(x)-\varphi_k(x)$是次数不高于$k-1$的多项式，根据定理5.6可知

$$(g_k - \varphi_k, g_k) = (g_k - \varphi_k, \varphi_k) = 0$$

因而

$$(g_k - \varphi_k, g_k - \varphi_k) = 0$$

因 $g_k(x),\varphi_k(x)\in C[a,b]$，故有

$$g_k(x)\equiv\varphi_k(x),\quad x\in[a,b]$$

证毕。

性质 3　设$\{\varphi_k(x)\}$是$[a,b]$上带权 $\rho(x)$的正交多项式系，则当 $k\geqslant 1$ 时，k 次正交多项式 $\varphi_k(x)$有 k 个互异实零点，并且全部位于开区间(a,b)内。

证　当 $k\geqslant 1$ 时，由于

$$\int_a^b\rho(x)\varphi_k(x)\varphi_0(x)\mathrm{d}x=0$$

且 $\varphi_0(x)$恒为非零常数，所以

$$\int_a^b\rho(x)\varphi_k(x)\mathrm{d}x=0$$

因而可知 $\varphi_k(x)$在(a,b)内必变号，$\varphi_k(x)$在(a,b)内必有奇重实零点。

设 $\varphi_k(x)$在(a,b)内总共有 m 个奇重实零点，记为 $\xi_1,\xi_2,\cdots,\xi_m$。假定 $m<k$。令

$$q(x)=(x-\xi_1)(x-\xi_2)\cdots(x-\xi_m)$$

则根据定理 5.6，必有

$$\int_a^b\rho(x)q(x)\varphi_k(x)\mathrm{d}x=0$$

但由于 $q(x)\varphi_k(x)$在(a,b)内无奇重实零点，因而 $q(x)\varphi_k(x)$在(a,b)内保持不变号，又显然 $q(x)\varphi_k(x)\not\equiv 0$，所以应有

$$\int_a^b p(x)q(x)\varphi_k(x)\mathrm{d}x\neq 0$$

所出现的矛盾证实 $m=k$。

证毕。

性质 4　设$\{\varphi_k(x)\}$是$[a,b]$上带权 $\rho(x)$的正交多项式系，则对于 $k\geqslant 1$，相邻三项有如下递推关系

$$\varphi_{k+1}(x)=\frac{a_{k+1}}{a_k}(x-\alpha_k)\varphi_k(x)-\frac{a_{k+1}a_{k-1}}{a_k^2}\lambda_{k-1}\varphi_{k-1}(x)\tag{5.73}$$

其中 a_k 是正交多项式 $\varphi_k(x)$的最高次项系数，并且

$$\alpha_k=\frac{(x\varphi_k,\varphi_k)}{(\varphi_k,\varphi_k)},\quad \lambda_{k-1}=\frac{(\varphi_k,\varphi_k)}{(\varphi_{k-1},\varphi_{k-1})}$$

证明从略。

在今后的叙述过程中，如果不指明所带的权函数 $\rho(x)$就表示 $\rho(x)\equiv 1$。

由于幂函数系$\{x^k(k=0,1,\cdots)\}$在任何区间上线性无关，所以，可采用 Gram - Schmidt(克莱姆-施密特)正交化方法由幂函数系产生在指定区间$[a,b]$上带指定权函数 $\rho(x)$的正交多项式系$\{\varphi_k(x)(k=0,1,\cdots)\}$，其中 $\varphi_k(x)$是最高次项系数为 1 的 k 次多项式。正交化方法如下：

$$\begin{cases}\varphi_0(x)\equiv 1\\ \varphi_{k+1}(x)=x^{k+1}-\displaystyle\sum_{j=0}^{k}a_{kj}\varphi_j(x)\quad(k=0,1,\cdots)\\ \qquad\text{其中 } a_{kj}=\dfrac{(x^{k+1},\varphi_j)}{(\varphi_j,\varphi_j)}\quad(j=0,1,\cdots,k)\end{cases}$$

例如，要构造区间$[-1,1]$上带权$\rho(x)=x^2$的正交多项式系$\{\varphi_k(x)(k=0,1,2,3)\}$，此时内积定义为

$$(f,g)=\int_{-1}^{1}x^2f(x)g(x)\mathrm{d}x$$

取$\varphi_0(x)\equiv 1$，令$\varphi_1(x)=x-a_{00}\varphi_0(x)$，由

$$a_{00}=\frac{(x,\varphi_0)}{(\varphi_0,\varphi_0)}=0$$

得

$$\varphi_1(x)=x$$

令$\varphi_2(x)=x^2-a_{10}\varphi_0(x)-a_{11}\varphi_1(x)$，由

$$a_{10}=\frac{(x^2,\varphi_0)}{(\varphi_0,\varphi_0)}=\frac{3}{5},\quad a_{11}=\frac{(x^2,\varphi_1)}{(\varphi_1,\varphi_1)}=0$$

得

$$\varphi_2(x)=x^2-\frac{3}{5}$$

令$\varphi_3(x)=x^3-a_{20}\varphi_0(x)-a_{21}\varphi_1(x)-a_{22}\varphi_2(x)$，由

$$a_{20}=\frac{(x^3,\varphi_0)}{(\varphi_0,\varphi_0)}=0,\quad a_{21}=\frac{(x^3,\varphi_1)}{(\varphi_1,\varphi_1)}=\frac{5}{7}$$

$$a_{22}=\frac{(x^3,\varphi_2)}{(\varphi_2,\varphi_2)}=0$$

得

$$\varphi_3(x)=x^3-\frac{5}{7}x$$

5.5.2 几种常用的正交多项式

1. Legendre 多项式

定义 由

$$\begin{cases}L_0(x)\equiv 1\\ L_n(x)=\dfrac{1}{2^n n!}\cdot\dfrac{\mathrm{d}^n}{\mathrm{d}x^n}[(x^2-1)^n]\quad(n=1,2,\cdots)\end{cases}\tag{5.74}$$

确定的$L_n(x)(n=0,1,\cdots)$称为Legendre(勒让德)多项式。

Legendre 多项式有以下重要性质：

(1) Legendre 多项式系$\{L_n(x)\}$是区间$[-1,1]$上的正交多项式系。

证 显然由式(5.74)确定的$L_n(x)$是x的n次多项式$(n=0,1,\cdots)$。考察积分

$$I_{mn}=\int_{-1}^{1}L_m(x)L_n(x)\mathrm{d}x=a\int_{-1}^{1}\frac{\mathrm{d}^m[(x^2-1)^m]}{\mathrm{d}x^m}\cdot\frac{\mathrm{d}^n[(x^2-1)^n]}{\mathrm{d}x^n}\mathrm{d}x$$

其中

$$a=\frac{1}{2^{m+n}m!n!}$$

当$m\neq n$时，不妨设$m<n$，连续使用m次分部积分公式，可得

$$I_{mn}=a\left\{\frac{\mathrm{d}^m}{\mathrm{d}x^m}[(x^2-1)^m]\cdot\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}[(x^2-1)^n]\Bigg|_{-1}^{1}-\right.$$
$$\left.\int_{-1}^{1}\frac{\mathrm{d}^{m+1}}{\mathrm{d}x^{m+1}}[(x^2-1)^m]\cdot\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}[(x^2-1)^n]\mathrm{d}x\right\}=$$

$$-a\int_{-1}^{1}\frac{\mathrm{d}^{m+1}}{\mathrm{d}x^{m+1}}[(x^2-1)^m]\cdot\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}[(x^2-1)^n]\mathrm{d}x=\cdots=$$

$$(-1)^m a\int_{-1}^{1}\frac{\mathrm{d}^{2m}}{\mathrm{d}x^{2m}}[(x^2-1)^m]\cdot\frac{\mathrm{d}^{n-m}}{\mathrm{d}x^{n-m}}[(x^2-1)^n]\mathrm{d}x=$$

$$(-1)^m a(2m)!\frac{\mathrm{d}^{n-m-1}}{\mathrm{d}x^{n-m-1}}[(x^2-1)^n]\bigg|_{-1}^{1}=0$$

当 $m=n$ 时，有

$$I_{mn}=(-1)^n\left(\frac{1}{2^n n!}\right)^2(2n)!\int_{-1}^{1}(x^2-1)^n\mathrm{d}x$$

令 $x=\sin t$，可算出

$$I_{mn}=\frac{(2n)!}{2^{2n-1}(n!)^2}\int_0^{\frac{\pi}{2}}\cos^{2n+1}t\mathrm{d}t=\frac{2}{2n+1}$$

于是有

$$\int_{-1}^{1}L_m(x)L_n(x)\mathrm{d}x=\begin{cases}0, & m\neq n\\ \dfrac{2}{2n+1}, & m=n\end{cases}$$

所以，$\{L_n(x)\}$ 是区间 $[-1,1]$ 上的正交多项式系。

证毕。

(2) $L_n(x)$ 的最高次项系数为

$$a_n=\frac{(2n)!}{2^n(n!)^2}$$

这个结论是明显的。

(3) n 为奇数时 $L_n(x)$ 为奇函数，n 为偶数时 $L_n(x)$ 为偶函数。

证　由

$$(x^2-1)^n=\sum_{j=0}^{n}(-1)^j\binom{n}{j}x^{2n-2j}$$

得

$$\frac{\mathrm{d}^n(x^2-1)^n}{\mathrm{d}x^n}=\sum_{j=0}^{m}(-1)^j\binom{n}{j}(2n-2j)(2n-2j-1)\cdots(2n-2j-n+1)x^{n-2j}$$

其中

$$m=\begin{cases}\dfrac{n}{2}, & \text{当 } n \text{ 为偶数}\\ \dfrac{n-1}{2}, & \text{当 } n \text{ 为奇数}\end{cases}$$

由此得知

$$L_n(-x)=(-1)^n L_n(x)$$

证毕。

(4) 满足递推关系：当 $n\geqslant 1$ 时，有

$$L_{n+1}(x)=\frac{2n+1}{n+1}xL_n(x)-\frac{n}{n+1}L_{n-1}(x)\tag{5.75}$$

证　由正交多项式性质(5.73)可推出此结论。

证毕。

由 $L_0(x)\equiv 1, L_1(x)=x$,利用式(5.75)就可推出

$$L_2(x) = \frac{1}{2}(3x^2 - 1)$$

$$L_3(x) = \frac{1}{2}(5x^3 - 3x)$$

$$L_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

$$L_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$$

$$\vdots$$

2. Chebyshev 多项式

定义 设 n 为非负整数,称

$$T_n(x) = \cos(n \arccos x), \quad -1 \leqslant x \leqslant 1 \tag{5.76}$$

为 Chebyshev(切比雪夫)多项式。

Chebyshev 多项式有以下的重要性质:

(1) $T_n(x)$是 x 的 n 次多项式,并且当 $n\geqslant 1$ 时,$T_n(x)$的最高次项系数为

$$a_n = 2^{n-1}$$

证 $T_0(x)$是 x 的 0 次多项式显然正确。为证明当 $n\geqslant 1$ 时 $T_n(x)$是 x 的 n 次多项式,并且最高次项系数为 2^{n-1},只须证明当$n\geqslant 1$时,三角恒等式

$$\cos n\theta = 2^{n-1}\cos^n\theta + \sum_{j=0}^{n-1} a_j^{(n)}\cos^j\theta \tag{5.77}$$

成立,其中 $a_j^{(n)}(j=0,1,\cdots,n-1)$为适当常数。今用归纳法证明之。当 $n=1$ 时,式(5.77)显然正确,其中 $a_0^{(1)}=0$。今设当 $n=1,2,\cdots,m$ 时式(5.77)正确,则当 $n=m+1$ 时,有

$$\begin{aligned}\cos(m+1)\theta = & 2\cos\theta\cos m\theta - \cos(m-1)\theta = \\ & 2\cos\theta\cdot\left(2^{m-1}\cos^m\theta + \sum_{j=0}^{m-1} a_j^{(m)}\cos^j\theta\right) - \\ & 2^{m-2}\cos^{m-1}\theta - \sum_{j=0}^{m-2} a_j^{(m-1)}\cos^j\theta = \\ & 2^m\cos^{m+1}\theta + \sum_{j=0}^{m} a_j^{(m+1)}\cos^j\theta\end{aligned}$$

按归纳法原理,对任何 $n\geqslant 1$,式(5.77)成立。令 $\theta=\arccos x$,代入式(5.77)得

$$T_n(x) = 2^{n-1}x^n + \sum_{j=0}^{n-1} a_j^{(n)}x^j, \quad -1 \leqslant x \leqslant 1$$

证毕。

(2) Chebyshev 多项式系$\{T_n(x)\}$是区间$[-1,1]$上带权$\frac{1}{\sqrt{1-x^2}}$的正交多项式系。

证 利用置换 $x=\cos\theta$,可得

$$\int_{-1}^{1}\frac{T_i(x)T_j(x)}{\sqrt{1-x^2}}\mathrm{d}x = \int_{\pi}^{0}\frac{\cos i\theta\cos j\theta}{\sin\theta}(-\sin\theta)\mathrm{d}\theta =$$

$$\int_0^{\pi} \cos i\theta \cos j\theta \mathrm{d}\theta = \begin{cases} 0, & i \neq j \\ \dfrac{\pi}{2}, & i = j \neq 0 \\ \pi, & i = j = 0 \end{cases}$$

证毕。

(3) $T_n(x)$满足递推关系

$$\begin{cases} T_0(x) \equiv 1 \\ T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad (n = 1,2,\cdots) \end{cases} \tag{5.78}$$

证　这个关系可由式(5.73)推出。但也可用下面方法证明之。

$T_0(x) \equiv 1$ 和 $T_1(x) = x$ 显然正确。由三角恒等式

$$\cos (n+1)\theta = 2\cos \theta \cos n\theta - \cos (n-1)\theta$$

令 $\theta = \arccos x$，即得

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

证毕。

由递推关系(5.78)，可得

$$\begin{aligned} T_0(x) &\equiv 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 2^2x^3 - 3x \\ T_4(x) &= 2^3x^4 - 8x^2 + 1 \\ T_5(x) &= 2^4x^5 - 20x^3 + 5x \\ &\vdots \end{aligned}$$

(4) 当 $n \geqslant 1$ 时，$T_n(x)$在开区间$(-1,1)$内有 n 个互异实零点，它们是

$$x_i = \cos \frac{2(n-i)+1}{2n}\pi \quad (i = 1,2,\cdots,n) \tag{5.79}$$

证　根据正交多项式性质 3，并且把式(5.79)直接代入式(5.76)即可知结论正确。

证毕。

(5) 当 n 为奇数时 $T_n(x)$是奇函数，当 n 为偶数时 $T_n(x)$为偶函数。

证　由三角恒等式

$$\arccos x + \arccos (-x) = \pi, \quad -1 \leqslant x \leqslant 1$$

可知

$$\begin{aligned} T_n(-x) &= \cos [n\arccos (-x)] = \cos [n(\pi - \arccos x)] = \\ &(-1)^n \cos (n\arccos x) = (-1)^n T_n(x) \end{aligned}$$

证毕。

3. Laguerre 多项式

定义　称

$$U_n(x) = \mathrm{e}^x \frac{\mathrm{d}^n (x^n \mathrm{e}^{-x})}{\mathrm{d}x^n} \quad (n = 0,1,\cdots)$$

为 Laguerre(拉盖尔)多项式。

Laguerre 多项式有以下的重要性质:

(1) $U_n(x)$是 x 的 n 次多项式,并且它的最高次项系数为 $a_n=(-1)^n$。

(2) Laguerre 多项式系$\{U_n(x)\}$是在区间$[0,\infty)$上带权 e^{-x}的正交多项式系。

事实上,有

$$\int_0^{\infty} e^{-x}U_m(x)U_n(x)dx = \begin{cases}0, & m\neq n\\ (n!)^2, & m=n\end{cases}$$

(3) $U_n(x)$满足递推关系

$$\begin{cases}U_0(x)\equiv 1\\ U_1(x)=-x+1\\ U_{n+1}(x)=(2n+1-x)U_n(x)-n^2U_{n-1}(x) \quad (n=1,2,\cdots)\end{cases}$$

4. Hermite 多项式

定义 称

$$H_n(x)=(-1)^n e^{x^2}\frac{d^n(e^{-x^2})}{dx^n} \quad (n=0,1,\cdots)$$

为 Hermite 多项式。

Hermite 多项式有以下的重要性质:

(1) $H_n(x)$是 x 的 n 次多项式,并且它的最高次项系数为$a_n=2^n$。

(2) Hermite 多项式系$\{H_n(x)\}$是在区间$(-\infty,\infty)$上带权 e^{-x^2} 的正交多项式系。

事实上有

$$\int_{-\infty}^{\infty} e^{-x^2}H_m(x)H_n(x)dx = \begin{cases}0, & m\neq n\\ 2^n n!\sqrt{\pi}, & m=n\end{cases}$$

(3) $H_n(x)$满足递推关系

$$\begin{cases}H_0(x)\equiv 1\\ H_1(x)=2x\\ H_{n+1}(x)=2xH_n(x)-2nH_{n-1}(x) \quad (n=1,2,\cdots)\end{cases}$$

5.6 函数的最佳平方逼近

5.6.1 最佳平方逼近的概念与解法

用简单函数 $p(x)$去近似一个给定在区间$[a,b]$上的连续函数 $f(x)$,是函数逼近要研究的问题。度量逼近误差的标准有许多种,本书只介绍其中重要的一种,其对应的函数逼近称为平方逼近,这是比较容易实现的一种逼近。

设函数组 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$都是$[a,b]$上的连续函数,并且在$[a,b]$上线性无关。以此函数组为基底,生成空间 $C[a,b]$的一个子空间

$$H_n=\text{Span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$$

则 H_n 中的任一个元素为

$$p(x) = \sum_{j=0}^{n} c_j \varphi_j(x)$$

对空间 $C[a,b]$ 中的任意两个函数 f 和 g，定义内积

$$(f,g) = \int_a^b \rho(x) f(x) g(x) \mathrm{d}x$$

其中 $\rho(x)$ 是 (a,b) 上的一个权函数。

定义　对于给定的函数 $f(x) \in C[a,b]$，若 $p^*(x) \in H_n$ 满足

$$(f - p^*, f - p^*) = \min_{p \in H_n} (f - p, f - p) \tag{5.80}$$

则称 $p^*(x)$ 为子空间 H_n 中对于 $f(x)$ 的最佳平方逼近元素。

特别地，如果 $\varphi_j(x) = x^j\ (j = 0,1,\cdots,n)$，则满足条件(5.80)的 $p^*(x) \in H_n$ 可称为函数 $f(x)$ 在区间 $[a,b]$ 上带权 $\rho(x)$ 的 n 次最佳平方逼近多项式。在具体问题中 $\rho(x)$ 是给定的，如果不指明 $\rho(x)$ 是什么就意味着 $\rho(x) \equiv 1$，这时内积定义为

$$(f,g) = \int_a^b f(x) g(x) \mathrm{d}x$$

定理 5.7　设 $f(x) \in C[a,b]$，$p^*(x) \in H_n$ 是子空间 H_n 中对于 $f(x)$ 的最佳平方逼近元素的充分必要条件是

$$(f - p^*, \varphi_j) = 0 \quad (j = 0,1,\cdots,n) \tag{5.81}$$

或对任一个 $p(x) \in H_n$，总有

$$(f - p^*, p) = 0$$

证

必要性　用反证法。设存在一个函数 $\varphi_k(x)$，使得

$$(f - p^*, \varphi_k) = \sigma_k \neq 0$$

令

$$q(x) = p^*(x) + \frac{\sigma_k}{(\varphi_k, \varphi_k)} \varphi_k(x)$$

显然 $q(x) \in H_n$，利用内积性质，可得

$$\begin{aligned}(f - q, f - q) &= (f - p^*, f - p^*) - \frac{2\sigma_k}{(\varphi_k, \varphi_k)}(f - p^*, \varphi_k) + \frac{\sigma_k^2}{(\varphi_k, \varphi_k)^2}(\varphi_k, \varphi_k) = \\ &(f - p^*, f - p^*) - \frac{\sigma_k^2}{(\varphi_k, \varphi_k)} < (f - p^*, f - p^*)\end{aligned}$$

这表示 $p^*(x)$ 不是最佳平方逼近元素，所出现的矛盾证实必要性成立。

充分性　设条件(5.81)成立。对任意的 $p(x) \in H_n$，有

$$\begin{aligned}(f - p, f - p) &= (f - p^* + p^* - p, f - p^* + p^* - p) = \\ &(f - p^*, f - p^*) + 2(f - p^*, p^* - p) + \\ &(p^* - p, p^* - p)\end{aligned}$$

由于

$$(f - p^*, p^* - p) = \sum_{j=0}^{n} (c_j^* - c_j)(f - p^*, \varphi_j) = 0$$

$$(p^* - p, p^* - p) \geq 0$$

所以有

$$(f-p,f-p) \geqslant (f-p^*,f-p^*)$$

因而 $p^*(x)$ 是 H_n 中对于 $f(x)$ 的最佳平方逼近元素。

证毕。

定理 5.8 设 $f(x)\in C[a,b]$，则在子空间 H_n 中对于 $f(x)$ 的最佳平方逼近元素是唯一的。

证 假定 $p(x)$ 和 $q(x)$ 都是 H_n 中对于 $f(x)$ 的最佳平方逼近元素，则由定理 5.7，有

$$(f-p,p-q)=(f-q,p-q)=0$$

因而有

$$(p-q,p-q)=(p-f+f-q,p-q)=$$
$$(p-f,p-q)+(f-q,p-q)=0$$

由此可知，在区间$[a,b]$上有

$$p(x)\equiv q(x)$$

证毕。

求最佳平方逼近元素 $p^*(x)=\sum_{k=0}^{n}c_k^*\varphi_k(x)$ 就是求它所含的系数 $c_k^*(k=0,1,\cdots,n)$。因

$$(f-p^*,\varphi_j)=(f,\varphi_j)-\sum_{k=0}^{n}c_k^*(\varphi_k,\varphi_j)$$

故条件(5.81)又可表达为

$$\sum_{k=0}^{n}c_k^*(\varphi_k,\varphi_j)=(f,\varphi_j)\quad(j=0,1,\cdots,n) \tag{5.82}$$

这是一个以 $c_0^*,c_1^*,\cdots,c_n^*$ 为未知数的 $n+1$ 元线性方程组。称式(5.82)为法方程或正规方程。它的系数矩阵为

$$\boldsymbol{G}=\begin{bmatrix}(\varphi_0,\varphi_0) & (\varphi_1,\varphi_0) & \cdots & (\varphi_n,\varphi_0)\\(\varphi_0,\varphi_1) & (\varphi_1,\varphi_1) & \cdots & (\varphi_n,\varphi_1)\\\vdots & \vdots & & \vdots\\(\varphi_0,\varphi_n) & (\varphi_1,\varphi_n) & \cdots & (\varphi_n,\varphi_n)\end{bmatrix}$$

由于 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$ 在区间$[a,b]$上连续且线性无关，因而矩阵 $\boldsymbol{G}$ 非奇异。如若不然，则齐次线性方程组

$$\sum_{k=0}^{n}\beta_k(\varphi_k,\varphi_j)=0\quad(j=0,1,\cdots,n)$$

有非零解 $(\beta_0,\beta_1,\cdots,\beta_n)^{\mathrm{T}}$，其中 $\beta_0,\beta_1,\cdots,\beta_n$ 不全为零。

令

$$\psi(x)=\sum_{k=0}^{n}\beta_k\varphi_k(x)$$

则有

$$(\psi,\varphi_j)=\sum_{k=0}^{n}\beta_k(\varphi_k,\varphi_j)=0\quad(j=0,1,\cdots,n)$$

因而有

$$(\psi,\psi)=\sum_{j=0}^{n}\beta_j(\psi,\varphi_j)=0$$

$$\psi(x) \equiv 0, \quad a \leqslant x \leqslant b$$

这与 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$ 在 $[a,b]$ 上线性无关相矛盾。由于 G 非奇异，故法方程(5.82)的解 $c_k^*(k=0,1,\cdots,n)$ 存在且唯一。

记 $\delta=(f-p^*,f-p^*)$，称 δ 为最佳平方逼近误差，$\sqrt{\delta}$ 称为均方误差。由于 $(f-p^*,p^*)=0$，所以有

$$\delta = (f-p^*,f) = (f,f)-(p^*,f) = (f,f)-\sum_{k=0}^{n} c_k^*(\varphi_k,f)$$

例 6　定义内积

$$(f,g) = \int_0^1 f(x)g(x)\mathrm{d}x$$

试在 $H_1=\mathrm{Span}\{1,x\}$ 中寻求对于 $f(x)=\sqrt{x}$ 的最佳平方逼近元素 $p(x)$。

解

$$\varphi_0(x)\equiv 1, \qquad \varphi_1(x)=x$$

$$(\varphi_0,\varphi_0)=\int_0^1 \mathrm{d}x = 1, \qquad (\varphi_1,\varphi_0)=\int_0^1 x\mathrm{d}x=\frac{1}{2}$$

$$(\varphi_1,\varphi_1)=\int_0^1 x^2\mathrm{d}x=\frac{1}{3}, \quad (\varphi_0,f)=\int_0^1\sqrt{x}\mathrm{d}x=\frac{2}{3}$$

$$(\varphi_1,f)=\int_0^1 x\sqrt{x}\mathrm{d}x=\frac{2}{5}$$

法方程为

$$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}\begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{2}{5} \end{bmatrix}$$

解得 $c_0=\frac{4}{15}, c_1=\frac{12}{15}$。所求的最佳平方逼近元素为

$$p(x)=\frac{4}{15}+\frac{12}{15}x, \quad 0\leqslant x\leqslant 1$$

5.6.2　正交函数系在最佳平方逼近中的应用

对于一般的基底 $\varphi_0,\varphi_1,\cdots,\varphi_n$，当 n 稍大时，计算法方程中的 (φ_k,φ_j) 以及求解法方程的计算量都是很大的。若采用 $1,x,x^2,\cdots,x^n$ 作基底，当 $\rho(x)\equiv 1$ 时，虽然 $(\varphi_k,\varphi_j)=(x^k,x^j)$ 容易计算，但由此形成的法方程系数矩阵 G 在 n 稍大时是病态矩阵，用单字长在计算机上求解法方程，其结果往往不太可靠。为避免上述的弊端，应采用正交基底。

设 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$ 是区间 $[a,b]$ 上带权 $\rho(x)$ 的正交函数系，即

$$(\varphi_k,\varphi_j)=\int_a^b \rho(x)\varphi_k(x)\varphi_j(x)\mathrm{d}x=0, \quad k\neq j$$

$$(\varphi_k,\varphi_k)>0$$

那么，法方程(5.82)的解为

$$c_k^*=\frac{(f,\varphi_k)}{(\varphi_k,\varphi_k)} \quad (k=0,1,\cdots,n) \tag{5.83}$$

1. Legendre 多项式的应用

对给定的函数 $f(x)\in C[-1,1]$，要求 $f(x)$ 在区间 $[-1,1]$ 上的 n 次最佳平方逼近多项式

$p_n(x)$。前已指出，这个问题相当于在内积为

$$(f,g)=\int_{-1}^{1}f(x)g(x)\mathrm{d}x$$

的情形下，在子空间

$$H_n=\mathrm{Span}\{1,x,x^2,\cdots,x^n\}$$

中寻求对 $f(x)$ 的最佳平方逼近元素 $p_n(x)$。今对该 H_n 另取一组基底，即

$$H_n=\mathrm{Span}\{L_0,L_1,\cdots,L_n\}$$

其中 $L_j(x)$ 是 j 次 Legendre 多项式。此时，法方程(5.82)的解可直接得到，就是

$$c_k^*=\frac{(f,L_k)}{(L_k,L_k)}=\frac{2k+1}{2}\int_{-1}^{1}L_k(x)f(x)\mathrm{d}x\quad(k=0,1,\cdots,n)\tag{5.84}$$

所求的 n 次最佳平方逼近多项式为

$$p_n(x)=\sum_{k=0}^{n}c_k^*L_k(x),\quad -1\leqslant x\leqslant 1\tag{5.85}$$

如果所给的区间不是$[-1,1]$，而是一般的有限区间$[a,b]$，那么，可以通过变量置换

$$x=\frac{a+b}{2}+\frac{b-a}{2}t$$

将它转化为区间 $-1\leqslant t\leqslant 1$ 上的情形来处理。

定理 5.9 设 $f(x)\in C[-1,1]$，则由式(5.85)和系数公式(5.84)所确定的多项式 $p_n(x)$，当 $n\to\infty$ 时均方收敛于 $f(x)$，即

$$\lim_{n\to\infty}(f-p_n,f-p_n)=0$$

若 $f''(x)\in C[-1,1]$，则当 $n\to\infty$ 时多项式 $p_n(x)$ 在区间$[-1,1]$上一致收敛于 $f(x)$，即

$$\lim_{n\to\infty}\max_{-1\leqslant x\leqslant 1}|f(x)-p_n(x)|=0$$

证明从略。

当 $n\to\infty$ 时由系数公式(5.84)所确定的式(5.85)就成为一个无穷级数：

$$\sum_{k=0}^{\infty}c_k^*L_k(x),\quad -1\leqslant x\leqslant 1$$

称此无穷级数为函数 $f(x)$ 的 Legendre 级数。

例 7 求 $f(x)=\mathrm{e}^x$ 在$[-1,1]$上的三次最佳平方逼近多项式。

解 采用 Legendre 多项式 $L_0(x),\cdots,L_3(x)$ 作为次数不高于 3 的多项式空间的基底。由

$$(f,L_0)=\int_{-1}^{1}\mathrm{e}^x\mathrm{d}x=2.350\,4$$

$$(f,L_1)=\int_{-1}^{1}x\mathrm{e}^x\mathrm{d}x=0.735\,8$$

$$(f,L_2)=\int_{-1}^{1}\left(\frac{3}{2}x^2-\frac{1}{2}\right)\mathrm{e}^x\mathrm{d}x=0.143\,1$$

$$(f,L_3)=\int_{-1}^{1}\left(\frac{5}{2}x^3-\frac{3}{2}x\right)\mathrm{e}^x\mathrm{d}x=0.020\,13$$

以及公式(5.84)，得

$$c_0^*=\frac{1}{2}(f,L_0)=1.175\,2,\quad c_1^*=\frac{3}{2}(f,L_1)=1.103\,6$$

$$c_2^*=\frac{5}{2}(f,L_2)=0.357\,8,\quad c_3^*=\frac{7}{2}(f,L_3)=0.070\,46$$

所求之最佳平方逼近多项式为

$$p_3(x)=1.175\,2L_0(x)+1.103\,6L_1(x)+0.357\,8L_2(x)+0.070\,46L_3(x)=$$
$$0.996\,3+0.997\,9x+0.536\,7x^2+0.176\,1x^3,\quad -1\leqslant x\leqslant 1$$

例 8　求 $f(x)=\sqrt{x}$ 在区间[0,1]上的一次最佳平方逼近多项式。

解　令 $x=\frac{1}{2}(1+t)$，则

$$f(x)=\frac{1}{\sqrt{2}}\sqrt{1+t}=\varphi(t),\quad -1\leqslant t\leqslant 1$$

先求 $\varphi(t)$ 在区间[-1,1]上的一次最佳平方逼近多项式 $q_1(t)$。由

$$c_0^*=\frac{1}{2}(\varphi,L_0)=\frac{1}{2}\int_{-1}^{1}\frac{1}{\sqrt{2}}\sqrt{1+t}\,\mathrm{d}t=\frac{2}{3}$$

$$c_1^*=\frac{3}{2}(\varphi,L_1)=\frac{3}{2}\int_{-1}^{1}\frac{t}{\sqrt{2}}\sqrt{1+t}\,\mathrm{d}t=\frac{6}{15}$$

可知

$$q_1(t)=\frac{2}{3}L_0(t)+\frac{6}{15}L_1(t)=\frac{2}{3}+\frac{6}{15}t,\quad -1\leqslant t\leqslant 1$$

把 $t=2x-1$ 代入 $q_1(t)$，就得 $\sqrt{x}$ 在区间[0,1]上的一次最佳平方逼近多项式

$$p_1(x)=\frac{2}{3}+\frac{6}{15}(2x-1)=\frac{4}{15}+\frac{12}{15}x,\quad 0\leqslant x\leqslant 1$$

2. Chebyshev 多项式的应用

定义内积

$$(f,g)=\int_{-1}^{1}\frac{f(x)g(x)}{\sqrt{1-x^2}}\mathrm{d}x$$

并取 $C[-1,1]$ 的一个子空间

$$H_n=\mathrm{Span}\{T_0,T_1,\cdots,T_n\}$$

其中 $T_j(x)$ 是 j 次 Chebyshev 多项式。H_n 中的任一元素为

$$p_n(x)=\frac{a_0}{2}+\sum_{j=1}^{n}a_jT_j(x),\quad -1\leqslant x\leqslant 1 \tag{5.86}$$

设 $f(x)\in C[-1,1]$。由于 $T_0,T_1,\cdots,T_n$ 是在区间[-1,1]上带权 $\frac{1}{\sqrt{1-x^2}}$ 的正交函数组，并且

$$(T_j,T_j)=\begin{cases}\pi, & j=0\\ \dfrac{\pi}{2}, & j\neq 0\end{cases}$$

所以，由式(5.83)可知，当

$$\begin{cases}a_0=\dfrac{2(f,T_0)}{(T_0,T_0)}=\dfrac{2}{\pi}\displaystyle\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x\\ a_j=\dfrac{(f,T_j)}{(T_j,T_j)}=\dfrac{2}{\pi}\displaystyle\int_{-1}^{1}\frac{f(x)T_j(x)}{\sqrt{1-x^2}}\mathrm{d}x\quad (j=1,2,\cdots,n)\end{cases} \tag{5.87}$$

时，式(5.86)所表示的 $p_n(x)$ 就是空间 H_n 中对于 $f(x)$ 的最佳平方逼近元素，也就是 $f(x)$ 在区间$[-1,1]$上带权 $\rho(x)=\dfrac{1}{\sqrt{1-x^2}}$ 的 n 次最佳平方逼近多项式。

定理 5.10 设 $f'(x)$ 在区间$[-1,1]$上存在且有界，那么由式(5.86)和系数公式(5.87)所确定的多项式 $p_n(x)$，当 $n\to\infty$ 时，在$[-1,1]$上一致收敛于函数 $f(x)$。

证明从略。

当 $n\to\infty$ 时，式(5.86)就成为一个无穷级数：

$$\frac{a_0}{2}+\sum_{j=0}^{\infty}a_jT_j(x),\quad -1\leqslant x\leqslant 1$$

称此无穷级数为函数 $f(x)$ 的 Chebyshev 级数，其中

$$a_j=\frac{2}{\pi}\int_{-1}^{1}\frac{f(x)T_j(x)}{\sqrt{1-x^2}}\mathrm{d}x\quad (j=0,1,2,\cdots)$$

由式(5.86)和(5.87)确定的多项式 $p_n(x)$ 亦称为函数 $f(x)$ 按 Chebyshev 多项式展开的部分和。

例 9 求函数 $f(x)=\arcsin x$ 按 Chebyshev 多项式展开的 $n=7$ 的部分和。

解

$$p_7(x)=\frac{a_0}{2}+\sum_{j=1}^{7}a_jT_j(x),\quad -1\leqslant x\leqslant 1$$

其中

$$a_{2k}=\frac{2}{\pi}\int_{-1}^{1}\frac{T_{2k}(x)\arcsin x}{\sqrt{1-x^2}}\mathrm{d}x=0\quad (k=0,1,2,3)$$

$$a_{2k-1}=\frac{2}{\pi}\int_{-1}^{1}\frac{T_{2k-1}(x)\arcsin x}{\sqrt{1-x^2}}\mathrm{d}x=$$

$$\frac{2}{\pi}\int_{0}^{\pi}\left(\frac{\pi}{2}-\theta\right)\cos\,(2k-1)\theta\mathrm{d}\theta=$$

$$\frac{4}{\pi}\,\frac{1}{(2k-1)^2}\quad (k=1,2,3,4)$$

即

$$p_7(x)=\frac{4}{\pi}\left[T_1(x)+\frac{T_3(x)}{9}+\frac{T_5(x)}{25}+\frac{T_7(x)}{49}\right]$$

把 $T_j(x)(j=1,3,5,7)$ 的表达式代入上式右端，得

$$p_7(x)=\frac{4}{\pi}\left(\frac{76}{105}x+\frac{248}{315}x^3-\frac{288}{175}x^5+\frac{64}{49}x^7\right),\quad -1\leqslant x\leqslant 1$$

设 $q_7(x)$ 是函数 $\arcsin x$ 的 Maclaurin 级数 $n=7$ 的部分和，那么，在区间$[-1,1]$上用例 9 所得的 $p_7(x)$ 近似代替 $\arcsin x$ 的精确度比用 $q_7(x)$ 高得多。原因是 Maclaurin 级数 $n=7$ 的部分和逼近 $\arcsin x$ 只在 $x=0$ 的近旁才有良好的精确度，而 Chebyshev 级数的部分和却是在区间$[-1,1]$上 $f(x)$ 的最佳平方逼近，其最佳逼近是对整个区间$[-1,1]$而言的。因此，函数的 Chebyshev 展开式常常用做函数在整个区间的近似计算，有最经济展开的称号。

3. 三角函数系的应用

当被逼近函数 $f(x)$ 是以 2π 为周期的函数时，宜用三角多项式作逼近函数。定义内积

$$(f,g)=\int_0^{2\pi}f(x)g(x)\mathrm{d}x$$

在空间

$$\mathscr{D}_n=\mathrm{Span}\{1,\cos x,\sin x,\cdots,\cos nx,\sin nx\}$$

中寻求对于 $f(x)$ 的最佳平方逼近元素

$$s_n(x)=\frac{a_0}{2}+\sum_{k=0}^{n}(a_k\cos kx+b_k\sin kx) \tag{5.88}$$

由于对 $k,j=0,1,\cdots,n$ 有

$$(\cos kx,\cos jx)=\begin{cases}0, & k\neq j\\ 2\pi, & k=j=0\\ \pi, & k=j\neq 0\end{cases}$$

$$(\sin kx,\sin jx)=\begin{cases}0, & k\neq j\\ \pi, & k=j\neq 0\end{cases}$$

$$(\cos kx,\sin jx)=0$$

故三角函数系 $\{1,\cos x,\sin x,\cdots,\cos nx,\sin nx\}$ 是区间 $[0,2\pi]$ 上的正交函数系。由式(5.83)可知，$f(x)$ 在 $[0,2\pi]$ 上的最佳平方逼近元素(5.88)中的系数为

$$\begin{cases}a_k=\dfrac{1}{\pi}\displaystyle\int_0^{2\pi}f(x)\cos kx\,\mathrm{d}x & (k=0,1,\cdots,n)\\ b_k=\dfrac{1}{\pi}\displaystyle\int_0^{2\pi}f(x)\sin kx\,\mathrm{d}x & (k=1,2,\cdots,n)\end{cases} \tag{5.89}$$

由公式(5.89)所表示的 a_k,b_k 称为 Fourier 系数。

由式(5.88)所表示的三角多项式 $s_n(x)$ 是 Fourier 级数

$$\frac{a_0}{2}+\sum_{k=1}^{\infty}(a_k\cos kx+b_k\sin kx)$$

的部分和。当 $f(x)\in C(-\infty,\infty)$ 且以 2π 为周期时，系数由公式(5.89)确定的 $s_n(x)$（当 $n\to\infty$ 时）在任意的 x 处收敛于 $f(x)$。

5.6.3 样条函数在最佳平方逼近中的应用

在区间 $[a,b]$ 上用分段 k 次多项式逼近函数 $f(x)$，又要求该分段 k 次多项式在 (a,b) 上具有一定的光滑性，这比在 $[a,b]$ 上用一个 k 次多项式逼近 $f(x)$ 的逼近精度要高。为此，可利用样条函数作为逼近函数。

问题的提法是：定义内积

$$(f,g)=\int_a^b f(x)g(x)\mathrm{d}x$$

给定区间 $[a,b]$ 上的一个等距分划

$$\pi:\quad x_i=a+ih\quad(i=0,1,\cdots,n)\quad h=\frac{b-a}{n}$$

在对应于分划 π 的 k 次样条函数空间 $\mathscr{D}_{k,\pi}$ 中寻求对于函数 $f(x)\in C[a,b]$ 的最佳平方逼近元素 $s(x)$。

为求解上述问题，首先选择空间 $\mathscr{D}_{k,\pi}$ 的基底。由于 π 是等距分划，故选择节点等距的 k 次 B 样条函数组

$$\varphi_j(x) = \Omega_k\left(\frac{x - x_{j-\frac{k-1}{2}}}{h}\right), \quad a \leqslant x \leqslant b \quad (j = 0,1,\cdots,n+k-1) \tag{5.90}$$

作空间 $\mathscr{D}_{k,\pi}$ 的基底。于是，所求的最佳平方逼近元素 $s(x)$ 为

$$s(x) = \sum_{j=0}^{n+k-1} c_j\varphi_j(x), \quad a \leqslant x \leqslant b$$

其中系数 $c_j(j=0,1,\cdots,n+k-1)$ 是法方程

$$\sum_{j=0}^{n+k-1} c_j(\varphi_j,\varphi_r) = (f,\varphi_r) \quad (r = 0,1,\cdots,n+k-1) \tag{5.91}$$

的解，而 $\varphi_j(x)$ 是由式(5.90)所表示的 k 次 B 样条。所得到的最佳平方逼近元素 $s(x)$ 是对应于分划 π 的 k 次样条，因而它是区间 $[a,b]$ 上的分段 k 次多项式且在 (a,b) 上有 $k-1$ 阶连续导数。

例 10 给定区间 $[-1,1]$ 上的分划

$$\pi: \quad x_0 = -1, \quad x_1 = 0, \quad x_2 = 1$$

在对应于分划 π 的三次样条函数空间 $\mathscr{D}_{3,\pi}$ 中寻求对于 $f(x)=\mathrm{e}^x$ 的最佳平方逼近元素 $s(x)$。

解 注意到本题 $n=2,k=3,x_{-1}=-2,x_3=2$，根据式(5.90)，可知

$$\mathscr{D}_{3,\pi} = \mathrm{Span}\{\Omega_3(x+2),\Omega_3(x+1),\Omega_3(x),\Omega_3(x-1),\Omega_3(x-2)\}$$

$$(\varphi_j,\varphi_r) = \int_{-1}^{1}\Omega_3(x+2-j)\Omega_3(x+2-r)\mathrm{d}x \quad (j,r = 0,1,2,3,4)$$

$$(f,\varphi_r) = \int_{-1}^{1}\mathrm{e}^x\Omega_3(x+2-r)\mathrm{d}x \quad (r = 0,1,2,3,4)$$

经积分计算，得

$$\begin{aligned}
&(\varphi_0,\varphi_0) = 0.003\,968\,25, && (\varphi_0,\varphi_1) = 0.025\,595\,2\\
&(\varphi_0,\varphi_2) = 0.011\,904\,8, && (\varphi_0,\varphi_3) = 0.000\,198\,413\\
&(\varphi_0,\varphi_4) = 0.0, && (\varphi_1,\varphi_1) = 0.239\,683\\
&(\varphi_1,\varphi_2) = 0.210\,714, && (\varphi_1,\varphi_3) = 0.023\,809\,5\\
&(\varphi_1,\varphi_4) = 0.000\,198\,413, && (\varphi_2,\varphi_2) = 0.471\,429\\
&(\varphi_2,\varphi_3) = 0.210\,714, && (\varphi_2,\varphi_4) = 0.011\,904\,8\\
&(\varphi_3,\varphi_3) = 0.239\,683, && (\varphi_3,\varphi_4) = 0.025\,592\,4\\
&(\varphi_4,\varphi_4) = 0.003\,968\,25, && (f,\varphi_0) = 0.018\,988\,2\\
&(f,\varphi_1) = 0.312\,426, && (f,\varphi_2) = 1.026\,73\\
&(f,\varphi_3) = 0.898\,349, && (f,\varphi_4) = 0.093\,906\,1
\end{aligned}$$

把上面的数据代入法方程(5.91)，并解之得

$$c_0 = 0.096\,718\,4, \quad c_1 = 0.316\,390, \quad c_2 = 0.843\,998$$
$$c_3 = 2.311\,82, \quad c_4 = 6.206\,92$$

所求的最佳平方逼近元素为

$$s(x) = \sum_{j=0}^{4} c_j\Omega_3(x+2-j) =$$
$$\begin{cases} 0.105\,379\,6x^3 + 0.470\,107x^2 + 0.997\,715x + 1.000\,70, & -1 \leqslant x \leqslant 0 \\ 0.247\,844x^3 + 0.470\,107x^2 + 0.997\,715x + 1.000\,70, & 0 < x \leqslant 1 \end{cases}$$

在区间[$-1,1$]上用此三次样条函数 $s(x)$ 逼近函数 e^x 的精度高于例 7 的三次最佳平方逼近多项式 $p_3(x)$。

5.6.4 曲线拟合与曲面拟合

1. 曲线拟合

设在 Oxy 直角坐标系中给定 $m+1$ 对数据(即坐标)

$$(x_i, y_i) \quad (i = 0,1,\cdots,m) \tag{5.92}$$

其中 $a=x_0<x_1<\cdots<x_m=b$。又选定 $n+1$ 个在区间[a,b]上连续且在点集{$x_i(i=0,1,\cdots,m)$}上线性无关的基函数 $\varphi_j(x)(j=0,1,\cdots,n)$,其中 $n\leqslant m$。问题是要在曲线族

$$y(x) = \sum_{j=0}^{n} c_j\varphi_j(x) \tag{5.93}$$

中寻找一曲线按照某种原则去拟合数据(5.92),用所得的拟合曲线去代替数据(5.92)所反映的函数关系。数据(5.92)中的 y_i 一般是在实验中通过测量得到的,总会带有观测误差,并且 m 往往很大,因此不能要求曲线 $y(x)$ 通过由数据(5.92)表示的所有点。

定义 若曲线

$$y^*(x) = \sum_{j=0}^{n} c_j^*\varphi_j(x) \tag{5.94}$$

使得

$$\sum_{i=0}^{m}\Big[\sum_{j=0}^{n} c_j^*\varphi_j(x_i) - y_i\Big]^2 = \min_{\{c_j\}}\sum_{i=0}^{m}\Big[\sum_{j=0}^{n} c_j\varphi_j(x_i) - y_i\Big]^2 \tag{5.95}$$

成立,则称曲线 $y^*(x)$ 为在曲线族(5.93)中按最小二乘原则确定的对于数据(5.92)的拟合曲线。

所谓最小二乘法求拟合曲线 $y^*(x)$,就是按条件(5.95)求出系数 $c_j^*(j=0,1,\cdots,n)$。记

$$\boldsymbol{\Phi}_j = (\varphi_j(x_0),\varphi_j(x_1),\cdots,\varphi_j(x_n))^{\mathrm{T}} \quad (j = 0,1,\cdots,n)$$
$$\boldsymbol{A} = [\boldsymbol{\Phi}_0,\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_n]$$
$$\boldsymbol{y} = (y_0,y_1,\cdots,y_m)^{\mathrm{T}}$$
$$\boldsymbol{c} = (c_0,c_1,\cdots,c_n)^{\mathrm{T}}$$

由于

$$\boldsymbol{Ac} = \sum_{j=0}^{n} c_j\boldsymbol{\Phi}_j = \Big(\sum_{j=0}^{n} c_j\varphi_j(x_0),\sum_{j=0}^{n} c_j\varphi_j(x_1),\cdots,\sum_{j=0}^{n} c_j\varphi_j(x_m)\Big)^{\mathrm{T}}$$

$$(\boldsymbol{Ac}-\boldsymbol{y},\boldsymbol{Ac}-\boldsymbol{y}) = \sum_{i=0}^{m}\Big[\sum_{j=0}^{n} c_j\varphi_j(x_i) - y_i\Big]^2$$

所以,所谓最小二乘法求拟合曲线 $y^*(x)$,就等价于求 $\boldsymbol{c}^*=(c_0^*,c_1^*,\cdots,c_n^*)^{\mathrm{T}}$,使得

$$(\boldsymbol{Ac}^*-\boldsymbol{y},\boldsymbol{Ac}^*-\boldsymbol{y}) = \min_{\boldsymbol{c}\in\mathbf{R}^{n+1}}(\boldsymbol{Ac}-\boldsymbol{y},\boldsymbol{Ac}-\boldsymbol{y}) \tag{5.96}$$

成立。

定理 5.11 设函数组{$\varphi_j(x)(j=0,1,\cdots,n)$}在点集{$x_i(i=0,1,\cdots,m)$}($n\leqslant m$)上线性无关,则 $\boldsymbol{c}^*\in\mathbf{R}^{n+1}$ 使得式(5.96)成立的充分必要条件是

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Ac}^* = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \tag{5.97}$$

证　必要性　记

$$F(\boldsymbol{c})=(\boldsymbol{Ac}-\boldsymbol{y},\boldsymbol{Ac}-\boldsymbol{y})=\left(\sum_{j=0}^{n}c_j\boldsymbol{\Phi}_j-\boldsymbol{y},\sum_{j=0}^{n}c_j\boldsymbol{\Phi}_j-\boldsymbol{y}\right)$$

因 $\boldsymbol{c}^*$ 是 $n+1$ 元函数 $F(\boldsymbol{c})$ 的极小点，故有

$$\left.\frac{\partial F}{\partial c_k}\right|_{c=c^*}=2\left(\boldsymbol{\Phi}_k,\sum_{j=0}^{n}c_j^*\boldsymbol{\Phi}_j-\boldsymbol{y}\right)=0\quad(k=0,1,\cdots,n)$$

即

$$\sum_{j=0}^{n}c_j^*(\boldsymbol{\Phi}_k,\boldsymbol{\Phi}_j)=(\boldsymbol{\Phi}_k,\boldsymbol{y})\quad(k=0,1,\cdots,n)$$

也即

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Ac}^*=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$$

充分性　设 $\boldsymbol{c}^*\in\mathbf{R}^{n+1}$ 已满足条件(5.97)。任取 $\boldsymbol{c}\in\mathbf{R}^{n+1}$，考察

$$\begin{aligned}F(\boldsymbol{c})=&(\boldsymbol{Ac}-\boldsymbol{y},\boldsymbol{Ac}-\boldsymbol{y})=\\&(\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*)+\boldsymbol{Ac}^*-\boldsymbol{y},\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*)+\boldsymbol{Ac}^*-\boldsymbol{y})=\\&(\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*),\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*))+2(\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*),\boldsymbol{Ac}^*-\boldsymbol{y})+\\&(\boldsymbol{Ac}^*-\boldsymbol{y},\boldsymbol{Ac}^*-\boldsymbol{y})\end{aligned}$$

由式(5.97)知

$$(\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*),\boldsymbol{Ac}^*-\boldsymbol{y})=(\boldsymbol{c}-\boldsymbol{c}^*)^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Ac}^*-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y})=0$$

又因向量组 $\{\boldsymbol{\Phi}_j(j=0,1,\cdots,n)\}$ 线性无关，故 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 是正定矩阵，因而当 $\boldsymbol{c}\neq\boldsymbol{c}^*$ 时有

$$(\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*),\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*))=(\boldsymbol{c}-\boldsymbol{c}^*)^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{c}-\boldsymbol{c}^*)>0$$

$$F(\boldsymbol{c})>F(\boldsymbol{c}^*)$$

即 $\boldsymbol{c}^*$ 使得式(5.96)成立。

证毕。

方程(5.97)称为关于曲线的最小二乘拟合的法方程或正规方程。由于 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 的正定性，故法方程(5.97)的解 $\boldsymbol{c}^*$ 存在且唯一。从法方程(5.97)中解出 $\boldsymbol{c}^*=(c_0^*,c_1^*,\cdots,c_n^*)^{\mathrm{T}}$ 就得到拟合曲线(5.94)。

拟合曲线 $y^*(x)$ 对数据(5.92)的拟合精度，可用误差平方和 σ 来描述，其中

$$\sigma=\sum_{i=0}^{m}[y^*(x_i)-y_i]^2$$

当 $m=n$ 时，由于 $\boldsymbol{\Phi}_0,\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_n$ 线性无关，故矩阵 $\boldsymbol{A}$ 是非奇异的方阵。此时，法方程(5.97)成为

$$\boldsymbol{Ac}^*=\boldsymbol{y}$$

这表明，拟合曲线 $y^*(x)=\sum_{j=0}^{n}c_j^*\varphi_j(x)$ 满足插值条件

$$\sum_{j=0}^{n}c_j^*\varphi_j(x_i)=y_i\quad(i=0,1,\cdots,m)$$

而成为插值曲线。

作曲线拟合，选择基函数是至关重要的，通常要根据具体问题的物理背景或坐标点 (x_i,y_i) $(i=0,1,\cdots,m)$ 的分布情况去选择。人们通常选择幂函数 $x^j(j=0,1,\cdots,n)$ 作基函数，这时，拟合曲线是 n 次多项式曲线 $y^*(x)=\sum_{j=0}^{n}c_j^*x^j$。但是，当 n 较大($n\geqslant7$)时，相应的法方程往往是病态的，n 越大病态越严重。为避免求解病态线性方程组，可构造在点集 $\{x_i(i=0,1,\cdots,$

m)}上正交的多项式系{$\varphi_j(x)(j=0,1,\cdots,n)$}作为基函数组，其中 $\varphi_j(x)$是 j 次多项式，且满足

$$(\boldsymbol{\Phi}_k,\boldsymbol{\Phi}_j)=\sum_{i=0}^{m}\varphi_k(x_i)\varphi_j(x_i)=\begin{cases}0, & k\neq j\\ a_j>0, & k=j\end{cases}\quad (k,j=0,1,\cdots,n;n\leqslant m)$$

此时，$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 是对角矩阵

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\mathrm{diag}((\boldsymbol{\Phi}_0,\boldsymbol{\Phi}_0),\cdots,(\boldsymbol{\Phi}_n,\boldsymbol{\Phi}_n))$$

法方程(5.97)的解为

$$c_j^*=\frac{(\boldsymbol{\Phi}_j,\boldsymbol{y})}{(\boldsymbol{\Phi}_j,\boldsymbol{\Phi}_j)}=\frac{\sum\limits_{i=0}^{m}y_i\varphi_j(x_i)}{\sum\limits_{i=0}^{m}\varphi_j^2(x_i)}\quad (j=0,1,\cdots,n)$$

拟合曲线 $y^*(x)=\sum\limits_{j=0}^{n}c_j^*\varphi_j(x)$ 也是 n 次多项式曲线。

在点集{$x_i(i=0,1,\cdots,m)$}上正交的多项式系{$\varphi_j(x)(j=0,1,\cdots,n;\ n\leqslant m)$}由下列递推公式构造：

$$\begin{cases}\varphi_0(x)\equiv 1\\ \varphi_1(x)=x-\alpha_0\\ \varphi_{j+1}(x)=(x-\alpha_j)\varphi_j(x)-\beta_j\varphi_{j-1}(x)\quad (j=1,2,\cdots,n-1)\end{cases}\tag{5.98}$$

其中

$$\alpha_j=\frac{\sum\limits_{i=0}^{m}x_i\varphi_j^2(x_i)}{\sum\limits_{i=0}^{m}\varphi_j^2(x_i)}\quad (j=0,1,\cdots,n-1)$$

$$\beta_j=\frac{\sum\limits_{i=0}^{m}\varphi_j^2(x_i)}{\sum\limits_{i=0}^{m}\varphi_{j-1}^2(x_i)}\quad (j=1,2,\cdots,n-1)$$

若 $n=m+1$，则按递推公式(5.98)构造出来的 $m+1$ 次多项式 $\varphi_{m+1}(x)$，必使得

$$\boldsymbol{\Phi}_{m+1}=(\varphi_{m+1}(x_0),\varphi_{m+1}(x_1),\cdots,\varphi_{m+1}(x_m))^{\mathrm{T}}=\boldsymbol{0}$$

例如，求在点集{1,2,3,4}上正交的多项式系{$\varphi_j(x)(j=0,1,2,3)$}。使用递推公式(5.98)可得

$$\begin{aligned}\varphi_0(x)&\equiv 1\\ \varphi_1(x)&=x-2.5\\ \varphi_2(x)&=x^2-5x+5\\ \varphi_3(x)&=x^3-7.5x^2+16.7x-10.5\end{aligned}$$

若继续求 α_3,β_3，可得 $\alpha_3=2.5,\beta_3=0.45$。

$$\varphi_4(x)=(x-\alpha_3)\varphi_3(x)-\beta_3\varphi_2(x)=x^4-10x^3+35x^2-50x+24$$

但此时 $\boldsymbol{\Phi}_4=(\varphi_4(1),\varphi_4(2),\varphi_4(3),\varphi_4(4))^{\mathrm{T}}=\boldsymbol{0}$。

设 $x_i(i=0,1,\cdots,m)$是 Chebyshev 多项式 $T_{m+1}(x)$的全部零点，即 $x_i=\cos\dfrac{2(m-i)+1}{2(m+1)}\pi$

$(i=0,1,\cdots,m)$。可以证明，Chebyshev 多项式系$\{T_j(x)(j=0,1,\cdots,n;\ n\leqslant m)\}$在点集$\{x_i(i=0,1,\cdots,m)\}$上正交。若记

$$\boldsymbol{\phi}_j=(T_j(x_0),T_j(x_1),\cdots,T_j(x_m))^{\mathrm{T}}\quad(j=0,1,\cdots,n)$$

则有

$$(\boldsymbol{\phi}_k,\boldsymbol{\phi}_j)=\begin{cases}0, & k\neq j\\ \frac{1}{2}(m+1), & k=j\neq 0\\ m+1, & k=j=0\end{cases}\quad(k,j=0,1,\cdots,n)$$

此等式称为 Chebyshev 多项式离散形式的正交性。

今以$\{T_j(x)(j=0,1,\cdots,n;n\leqslant m)\}$为基函数组对数据$(x_i,y_i)(i=0,1,\cdots,m)$进行最小二乘拟合，其中$x_i(i=0,1,\cdots,m)$是$T_{m+1}(x)$的全部零点，那么，拟合曲线为

$$y^*(x)=\frac{c_0^*}{2}+\sum_{j=1}^{n}c_j^*T_j(x)$$

其中

$$c_j^*=\frac{(\boldsymbol{\phi}_j,\boldsymbol{y})}{(\boldsymbol{\phi}_j,\boldsymbol{\phi}_j)}=\frac{2}{m+1}\sum_{i=0}^{m}y_iT_j(x_i)\quad(j=0,1,\cdots,n)$$

若$x_i(i=0,1,\cdots,m)$只能分布在区间$[a,b]$内，并可预先设定，则在采集数据时可取x_i为

$$x_i=\frac{a+b}{2}+\frac{b-a}{2}t_i\quad(i=0,1,\cdots,m)$$

其中

$$t_i=\cos\frac{2(m-i)+1}{2(m+1)}\pi\quad(i=0,1,\cdots,m)$$

是$m+1$次 Chebyshev 多项式$T_{m+1}(t)$的全部零点。于是，函数组

$$\tilde{T}_j(x)=T_j\left(\frac{1}{b-a}(2x-a-b)\right)\quad(j=0,1,\cdots,n;n\leqslant m)$$

必在所设定的点集$\{x_i(i=0,1,\cdots,m)\}$上正交；并且，当用曲线

$$y^*(x)=\frac{c_0^*}{2}+\sum_{j=1}^{n}c_j^*\tilde{T}_j(x)\tag{5.99}$$

对数据$(x_i,y_i)(i=0,1,\cdots,m)$进行最小二乘拟合时，式(5.99)中的系数$c_j^*$的计算公式为

$$c_j^*=\frac{2}{m+1}\sum_{i=0}^{m}y_i\tilde{T}_j(x_i)=\frac{2}{m+1}\sum_{i=0}^{m}y_iT_j(t_i)\quad(j=0,1,\cdots,n)$$

曲线族(5.93)的函数结构是某个基函数组的线性组合，即$y(x)$关于待定参数$c_0,c_1,\cdots,c_n$是线性的。用这种函数形式作曲线的最小二乘拟合称为线性最小二乘问题。如果曲线族的函数结构$y(x)=f(x,c_0,c_1,\cdots,c_n)$关于待定参数$c_0,c_1,\cdots,c_n$是非线性的，则最小二乘问题

$$\sum_{i=0}^{m}[f(x_i,c_0,c_1,\cdots,c_n)-y_i]^2=\min,\quad n\leqslant m$$

称为非线性最小二乘问题。本书不讨论此问题的一般求解方法。但是，有一些非线性最小二乘问题是可以化为线性最小二乘问题来求解的。例如，曲线族$y(x)=\dfrac{1}{c_0+c_1x}$可化为$u=c_0+c_1x$，其中$u=\dfrac{1}{y}$；曲线族$y(x)=ae^{bx}$可化为$u=c_0+c_1x$，其中$u=\ln y,c_0=\ln a,c_1=b$。

例 11　给定数据表

x	-2	-1	0	1	2
y	-0.1	0.1	0.4	0.9	1.6

试分别用二次和三次多项式以最小二乘法拟合所给数据，并比较其优劣。

解　(1)
$$y(x)=c_0+c_1x+c_2x^2$$

$$\boldsymbol{A}=\begin{bmatrix}1 & -2 & 4\\1 & -1 & 1\\1 & 0 & 0\\1 & 1 & 1\\1 & 2 & 4\end{bmatrix},\quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\begin{bmatrix}5 & 0 & 10\\0 & 10 & 0\\10 & 0 & 34\end{bmatrix},\quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}=\begin{bmatrix}2.9\\4.2\\7\end{bmatrix}$$

法方程
$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{c}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$$
的解为
$$c_0=0.408\,6,\quad c_1=0.42,\quad c_2=0.085\,7$$
所求二次多项式为
$$y(x)=0.408\,6+0.42x+0.085\,7x^2$$
误差平方和为
$$\sigma_2=0.001\,16$$

(2)
$$y(x)=c_0+c_1x+c_2x^2+c_3x^3$$

$$\boldsymbol{A}=\begin{bmatrix}1 & -2 & 4 & -8\\1 & -1 & 1 & -1\\1 & 0 & 0 & 0\\1 & 1 & 1 & 1\\1 & 2 & 4 & 8\end{bmatrix},\quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\begin{bmatrix}5 & 0 & 10 & 0\\0 & 10 & 0 & 34\\10 & 0 & 34 & 0\\0 & 34 & 0 & 130\end{bmatrix}$$
$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}=(2.9,4.2,7,14.4)^{\mathrm{T}}$$

法方程
$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{c}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$$
的解为
$$c_0=0.408\,6,\quad c_1=0.391\,67,\quad c_2=0.085\,7,\quad c_3=0.008\,33$$
得到三次多项式
$$y(x)=0.408\,6+0.391\,67x+0.085\,7x^2+0.008\,33x^3$$
误差平方和为
$$\sigma_3=0.000\,194$$
由于 $\sigma_3<\sigma_2$，所以，用三次多项式拟合所给数据优于用二次多项式去拟合。

例 12　已知一组实验数据，见表 5－3。

表 5-3 例 12 实验数据

i	x_i	y_i	i	x_i	y_i
0	2	106.42	6	11	110.59
1	3	108.20	7	14	110.60
2	4	109.50	8	16	110.76
3	7	110.00	9	18	111.00
4	8	109.93	10	19	111.20
5	10	110.49			

试以最小二乘原则求一个函数拟合表 5-3 的数据。

解 在 Oxy 坐标平面上描出各点 $(x_i, y_i)(i=0,1,\cdots,10)$，并大致描出曲线，见图 5-6。

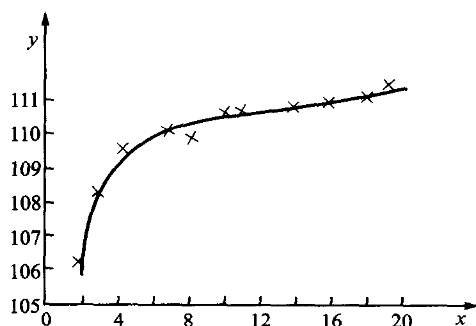


图 5-6 拟合数据的曲线示意图

凭观察分析这条曲线近似于一个什么类型的函数。这里，选择两种类型的函数。

第一种，选择双曲线型的函数：

$$y(x) = c_0 + \frac{c_1}{x}$$

这时，$\varphi_0(x) \equiv 1, \varphi_1(x) = \dfrac{1}{x}$。

$$\boldsymbol{\Phi}_0 = (1,1,\cdots,1)^{\mathrm{T}}$$

$$\boldsymbol{\Phi}_1 = \left(\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{7},\frac{1}{8},\frac{1}{10},\frac{1}{11},\frac{1}{14},\frac{1}{16},\frac{1}{18},\frac{1}{19}\right)^{\mathrm{T}}$$

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{\Phi}_0 & \boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{\Phi}_1 \\ \boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{\Phi}_0 & \boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{\Phi}_1 \end{bmatrix} = \begin{bmatrix} 11 & 1.784\,2 \\ 1.784\,2 & 0.492\,77 \end{bmatrix}$$

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} \boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} 1\,208.69 \\ 194.052 \end{bmatrix}$$

法方程 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{c} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$ 的解为

$$c_0 = 111.476, \quad c_1 = -9.832\,06$$

得双曲线函数

$$y(x) = 111.476 - \frac{9.832\,06}{x}$$

它对于 $y_i(i=0,1,\cdots,10)$ 的误差平方和为

$$\sigma_1 = 0.461\,3$$

第二种，选择指数类型的函数：

$$y(x) = ae^{\frac{b}{x}}$$

这时，$y(x)$不是某组已知函数的线性组合。对上式两边取对数，得

$$\ln y(x) = \ln a + \frac{b}{x}$$

记 $u=\ln y, c_0=\ln a, c_1=b$，则有

$$u(x) = c_0 + \frac{c_1}{x}$$

把原(x_i, y_i)数据表换成(x_i, u_i)数据表，见表 5-4。

表 5-4　例 12 数据换算表

i	x_i	$u_i=\ln y_i$	i	x_i	$u_i=\ln y_i$
0	2	4.667 39	6	11	4.705 83
1	3	4.683 98	7	14	4.705 92
2	4	4.695 92	8	16	4.707 37
3	7	4.700 48	9	18	4.709 53
4	8	4.699 84	10	19	4.711 33
5	10	4.704 93			

法方程的系数矩阵 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 与前面第一种的相同，而法方程的右端向量为

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} = \begin{bmatrix} \boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{u} \\ \boldsymbol{\Phi}_1^{\mathrm{T}}\boldsymbol{u} \end{bmatrix} = \begin{bmatrix} 51.692\,5 \\ 8.366\,23 \end{bmatrix}$$

这里 $\boldsymbol{u}=(u_0, u_1, \cdots, u_{10})^{\mathrm{T}}$。求解

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{c} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}$$

得 $c_0=4.714\,0, c_1=-0.090\,321$，由此得

$$a = e^{c_0} = 111.494, \quad b = c_1 = -0.090\,321$$

所求的指数函数为

$$y(x) = 111.494e^{-\frac{0.090\,321}{x}}$$

它对于数据 $y_i(i=0,1,\cdots,10)$的误差平方和为

$$\sigma_2 = 0.471\,9$$

2. 曲面拟合

设在三维直角坐标系 $Oxyu$ 中给定$(m+1)\times(n+1)$个点（即三维坐标）

$$(x_i, y_j, u_{ij}) \quad (i = 0,1,\cdots,m; j = 0,1,\cdots,n) \tag{5.100}$$

其中 $a=x_0<x_1<\cdots<x_m=b, c=y_0<y_1<\cdots<y_n=d$。选定 $M+1$ 个 x 的函数$\{\varphi_r(x)(r=0,1,\cdots,M)\}(M<m)$以及 $N+1$ 个 y 的函数$\{\psi_s(y)(s=0,1,\cdots,N)\}(N<n)$。这两个函数组分别在区间$[a,b]$和区间$[c,d]$上连续，且分别在点集$\{x_i(i=0,1,\cdots,m)\}$和点集$\{y_j(j=0,1,\cdots,n)\}$上线性无关。以函数组

$$\{\varphi_r(x)\psi_s(y)(r = 0,1,\cdots,M; s = 0,1,\cdots,N)\}$$

为基函数，称为乘积型基函数，构成以$\{c_{rs}\}$为参数的曲面族

$$p(x,y) = \sum_{s=0}^{N}\sum_{r=0}^{M} c_{rs}\varphi_r(x)\psi_s(y) \tag{5.101}$$

定义 若参数$\{c_{rs}^*\}$使得

$$\sum_{j=0}^{n}\sum_{i=0}^{m}\Big[\sum_{s=0}^{N}\sum_{r=0}^{M}c_{rs}^*\varphi_r(x_i)\psi_s(y_j)-u_{ij}\Big]^2=\min \tag{5.102}$$

成立，则称相应的曲面$p^*(x,y)$为在曲面族(5.101)中按最小二乘原则确定的对于数据(5.100)的拟合曲面。

这种拟合方法又称为乘积型最小二乘法，它属于二维情形的线性最小二乘问题。可分解为对x和y作曲线的最小二乘拟合两个步骤来实现乘积型最小二乘法，并由此推出拟合曲面参数$\{c_{rs}^*\}$的计算公式。

第一步，固定y_j，以$\{\varphi_r(x)\}$为基函数组对数据$(x_i,u_{ij})(i=0,1,\cdots,m)$作最小二乘拟合，得到$n+1$条拟合曲线

$$q_j(x)=\sum_{r=0}^{M}\alpha_{rj}\varphi_r(x)\quad(j=0,1,\cdots,n) \tag{5.103}$$

其中$(\alpha_{0j},\alpha_{1j},\cdots,\alpha_{Mj})^{\mathrm{T}}=\boldsymbol{\alpha}_j$是法方程

$$\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\alpha}_j=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_j\quad(j=0,1,\cdots,n) \tag{5.104}$$

的解，而$\boldsymbol{B}=[\varphi_r(x_i)]_{(m+1)\times(M+1)}$，$\boldsymbol{u}_j=(u_{0j},u_{1j},\cdots,u_{mj})^{\mathrm{T}}$。记$\boldsymbol{A}=[\alpha_{rj}]_{(M+1)\times(n+1)}$，$\boldsymbol{U}=[u_{ij}]_{(m+1)\times(n+1)}$，则由式(5.104)可知

$$\boldsymbol{A}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U} \tag{5.105}$$

第二步，任意固定x，以$\{\psi_s(y)\}$为基函数组对数据$(y_j,q_j(x))(j=0,1,\cdots,n)$作最小二乘拟合得到所求的拟合曲面

$$p^*(x,y)=\sum_{s=0}^{N}\beta_s(x)\psi_s(y) \tag{5.106}$$

其中$(\beta_0(x),\beta_1(x),\cdots,\beta_N(x))^{\mathrm{T}}=\boldsymbol{\beta}(x)$是法方程

$$\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{\beta}(x)=\boldsymbol{G}^{\mathrm{T}}\boldsymbol{q}(x) \tag{5.107}$$

的解，而$\boldsymbol{G}=[\psi_s(y_j)]_{(n+1)\times(N+1)}$，$\boldsymbol{q}(x)=(q_0(x),q_1(x),\cdots,q_n(x))^{\mathrm{T}}$。

拟合曲面(5.106)的表达式与表达式(5.101)在形式上是一致的。设$(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}=[\gamma_{sj}]_{(N+1)\times(n+1)}$，则由式(5.107)和式(5.103)可知

$$\beta_s(x)=\sum_{j=0}^{n}\gamma_{sj}\,q_j(x)=\sum_{r=0}^{M}\sum_{j=0}^{n}\alpha_{rj}\gamma_{sj}\varphi_r(x)$$

将上式代入式(5.106)，得到与式(5.101)形式一致的拟合曲面

$$p^*(x,y)=\sum_{s=0}^{N}\sum_{r=0}^{M}c_{rs}^*\varphi_r(x)\psi_s(y) \tag{5.108}$$

其中

$$c_{rs}^*=\sum_{j=0}^{n}\alpha_{rj}\gamma_{sj}\quad(r=0,1,\cdots,M;s=0,1,\cdots,N) \tag{5.109}$$

记$\boldsymbol{C}=[c_{rs}^*]_{(M+1)\times(N+1)}$，则系数$c_{rs}^*$的表达式(5.109)可表示为矩阵形式

$$\boldsymbol{C}=\boldsymbol{A}[(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}]^{\mathrm{T}}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{G}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1} \tag{5.110}$$

只须按照公式(5.110)计算系数$c_{rs}^*(r=0,1,\cdots,M;s=0,1,\cdots,N)$，即可得出要求的拟合曲面(5.108)。

最后还应验证，由公式(5.110)确定的系数$\{c_{rs}^*\}$确实使条件(5.102)成立。此问题留给读者思考。

曲面(5.108)对数据(5.100)的拟合精度就用误差平方和

$$\sigma = \sum_{j=0}^{n}\sum_{i=0}^{m}[p^*(x_i,y_j)-u_{ij}]^2$$

来描述。

例 13　给定表 5-5 的数据

表 5-5　例 13 数据表

x \ y (u)	−2	−1	0	1	2	3
−1	6	3	2	3	6	11
−0.5	4.4	1.51	0.5	1.48	4.6	9.4
0	4	1	0	1	4	9
0.5	4.6	1.49	0.5	1.52	4.4	9.6
1	6	3	2	3	6	11
1.5	8.4	5.6	4.5	5.4	8.6	13.4

试取 $\varphi_0(x)\equiv 1,\varphi_1(x)=x^2,\psi_s(y)=y^s(s=0,1,2)$，对所给数据作乘积型最小二乘拟合。

解

$$\boldsymbol{B}=[\varphi_r(x_i)]_{6\times 2}=\begin{bmatrix}1 & 1\\ 1 & 0.25\\ 1 & 0\\ 1 & 0.25\\ 1 & 1\\ 1 & 2.25\end{bmatrix}$$

$$\boldsymbol{G}=[\psi_s(y_j)]_{6\times 3}=\begin{bmatrix}1 & -2 & 4\\ 1 & -1 & 1\\ 1 & 0 & 0\\ 1 & 1 & 1\\ 1 & 2 & 4\\ 1 & 3 & 9\end{bmatrix}$$

$\boldsymbol{U}$ 就是表 5-5 中的数据$[u_{ij}]_{6\times 6}$。求出$(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}$和$(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}$之后，代入公式(5.110)，得

$$\boldsymbol{C}=\begin{bmatrix}-0.003\,890\,6 & -0.001\,945\,3 & 1.002\,4\\ 2.009\,7 & 0.004\,863\,3 & -0.006\,079\,1\end{bmatrix}$$

所求的拟合曲面为

$$\begin{aligned}p^*(x,y)=&c_{00}+c_{10}x^2+c_{01}y+c_{11}x^2y+c_{02}y^2+c_{12}x^2y^2=\\ &-0.003\,890\,6+2.009\,7x^2-0.001\,945\,3y+\\ &0.004\,863\,3x^2y+1.002\,4y^2-0.006\,079\,1x^2y^2\end{aligned}$$

其拟合精度为 $\sigma=\sum_{j=0}^{5}\sum_{i=0}^{5}[p^*(x_i,y_j)-u_{ij}]^2=0.103\,58$。

习　　题

1. 证明：函数系$\{x^{j-1}(j=1,2,\cdots,n+1)\}$在任何点集 $X=\{x_1,x_2,\cdots,x_m\}$上线性无关，其

中 $x_1,x_2,\cdots,x_m$ 互异，$m\geqslant n+1$。

2. 证明：函数系 $\left\{\frac{1}{x^{j-1}}(j=1,2,\cdots,n+1)\right\}$ 在任何点集 $X=\{x_1,x_2,\cdots,x_m\}$ 上线性无关，其中 $x_1,x_2,\cdots,x_m$ 大于零且互异，$m\geqslant n+1$。

3. 证明：函数系 $\{\sin x,\sin 2x,\sin 3x\}$ 在点集 $X=\left\{\frac{\pi}{4}i(i=0,1,\cdots,8)\right\}$ 上线性无关。

4. 给定数表

x	10	11	12	13	14
$f(x)$	210	230	240	235	230

试建立 $f(x)$ 的四次 Lagrange 插值多项式。

5. 设 $l_k(x)(k=0,1,\cdots,n)$ 是以互异的 $x_0,x_1,\cdots,x_n$ 为节点的 Lagrange 插值基函数，证明：

(1) $\sum\limits_{k=0}^{n} x_k^m l_k(x)\equiv x^m \quad (m=0,1,\cdots,n)$；

(2) $\sum\limits_{k=0}^{n} (x_k-x)^m l_k(x)\equiv 0 \quad (m=1,2,\cdots,n)$。

6. 证明：若 $f(x)$ 是次数不超过 n 的多项式，那么任取 $n+1$ 个实数为节点所作的 $f(x)$ 的 n 次插值多项式就一定是 $f(x)$ 自身。

7. 给定数表

x	0.10	0.15	0.25	0.30
e^{-x}	0.904 837	0.860 708	0.778 801	0.740 818

(1) 以 0.15,0.25,0.30 为节点作二次插值计算 $e^{-0.23}$ 的近似值，并估计截断误差。

(2) 以 0.10,0.15 为节点作线性插值计算 $e^{-0.14}$ 的近似值，并估计截断误差。

8. 给定数表

x	−2	−1.5	0.5	1	1.5
$f(x)$	21	23	22	21	20

试建立差商表，并写出 $f(x)$ 的四次 Newton 插值多项式。

9. 设实数 $x_0,x_1,\cdots,x_n$ 互异，证明差商具有下列性质：

(1) 若 $F(x)=cf(x)$，则

$$F[x_0,x_1,\cdots,x_n]=cf[x_0,x_1,\cdots,x_n]$$

(2) 若 $F(x)=f(x)+g(x)$，则

$$F[x_0,x_1,\cdots,x_n]=f[x_0,x_1,\cdots,x_n]+g[x_0,x_1,\cdots,x_n]$$

10. 设 $f(x)=\sum\limits_{j=0}^{n} a_j x^j$，实数 $x_0,x_1,\cdots,x_m(m>n)$ 互异，试计算 $f[x_0,x_1,\cdots,x_k](k=n,n+1,\cdots,m)$ 的值。

11. 证明：若 $f(x)=\sum\limits_{j=0}^{n} a_j x^j$ 且 $a_n\neq 0$，则 $f[x_0,x](x\neq x_0)$ 是 x 的 $n-1$ 次多项式。

12. 给定数表 $(x_i,f(x_i))(i=0,1,\cdots,n)$，其中 $x_i=a+ih\ (i=0,1,\cdots,n;h>0)$。设 $x_k<$

$x<x_k+\frac{h}{2}(1\leqslant k\leqslant n-2)$，要用：(1) 线性插值；(2) 二次插值；(3) 三次插值计算 $f(x)$ 的近似值，试分别写出所用的插值公式及其余项。

13. 试利用第 8 题的差商表，构造两个二次 Newton 插值多项式以分别计算 $f(-1)$ 和 $f(0.8)$ 的近似值，又构造一个三次 Newton 插值多项式以计算 $f(0)$ 的近似值。

14. 设将 $\sin x$ 在区间 $\left[0,\frac{\pi}{2}\right]$ 上的值列成节点等距分布的函数表，要求用二次插值计算非节点处的 $\sin x$ 值的截断误差不超过 10^{-6}，问节点的步长 h 最大不超过多少？

15. 给定数表

x	-0.1	0.3	0.7	1.1	1.5
$f(x)$	0.995	0.955	0.765	0.454	0.100

若不计舍入误差，那么，用分段二次插值计算 $f(x)(-0.1\leqslant x\leqslant 1.5)$ 的近似值能保证有多少位有效数字？其中已知 $\max\limits_{-0.1\leqslant x\leqslant 1.5}|f'''(x)|=1$。

16. 给定函数 $u=f(x,y)$ 的一个数表（见表 5-6），试分别采用：(1) 双一次插值；(2) 对 x 二次、对 y 一次的二元插值；(3) 双二次插值计算 $f(0.8,1.25)$ 的近似值。

表 5-6　习题 16 数据表

u \ x / y	0	0.5	1	1.5	2
1	10	12	13	15	14
1.2	14	15	17	18	16
1.4	15	16	18	17	15
1.6	13	14	16	15	13

17. 给定函数表

x	0	1	2
$f(x)$	1	2	1
$f'(x)$		0	-1

试构造一个次数不高于 4 的多项式 $H_4(x)$，使其满足条件：

$$H_4(0)=f(0),\quad H_4(1)=f(1),\quad H_4(2)=f(2)$$
$$H'_4(1)=f'(1),\quad H'_4(2)=f'(2)$$

并写出余项 $f(x)-H_4(x)$ 的表达式[设 $f(x)$ 处处有五阶导数]。

18. 设 x_0,x_1 互异，$f(x)$ 处处有三阶导数，试构造一个次数不高于 2 的多项式 $H_2(x)$，使其满足条件：

$$H_2(x_i)=f(x_i)\quad(i=0,1),\quad H'_2(x_0)=f'(x_0)$$

并写出余项 $f(x)-H_2(x)$ 的表达式。

19. 给定数表

x	0	1.5
$f(x)$	1	0.070 7
$f'(x)$	0	-0.997

试构造一个次数不高于 3 的多项式 $H_3(x)$，使其满足条件：

$$H_3(0)=f(0),\quad H_3(1.5)=f(1.5)$$
$$H'_3(0)=f'(0),\quad H'_3(1.5)=f'(1.5)$$

设已知 $\max\limits_{0\leqslant x\leqslant 1.5}|f^{(4)}(x)|\leqslant 1$，试估计误差 $\max\limits_{0\leqslant x\leqslant 1.5}|f(x)-H_3(x)|$。

20. 对任意正整数 k，证明：

$$x_+^k+(-1)^k(-x)_+^k=x^k$$

21. 试证：函数 x_+^m（$m>1$ 是整数）在区间 $(-\infty,\infty)$ 上有 $m-1$ 阶连续导数，并且在 $x=0$ 处不存在 m 阶导数。

22. 试判断下列两个函数是否为定义在区间 $[-1,1]$ 上以 $x_1=0$ 为内节点的三次样条函数，若是，则把它写成表达式(5.25)的形式：

(1) $s(x)=\begin{cases}-x^3-3x^2-x+2, & -1\leqslant x\leqslant 0\\ x^3-3x^2-x+2, & 0<x\leqslant 1\end{cases}$；

(2) $s(x)=\begin{cases}5x^3+3x^2-x+2, & -1\leqslant x\leqslant 0\\ x^3-3x^2-x+2, & 0<x\leqslant 1\end{cases}$。

23. 试验证下列函数是定义在区间 $[0,3]$ 上以 $x_1=1, x_2=2$ 为内节点的三次样条函数，并把它写成表达式(5.25)的形式：

$$s(x)=\begin{cases}1-x^3, & 0\leqslant x<1\\ 3x-3x^2, & 1<x\leqslant 2\\ 16-21x+9x^2-2x^3, & 2<x\leqslant 3\end{cases}$$

24. 试写出三次样条函数 $\Omega_3(2(x-1))$ 的分段表达式，并画出它的图形。

25. 给定数表

x	0	1	2	3
$f(x)$	3	6	8	7
$f''(x)$	-2			-1

利用三次 B 样条，求以 $x_0=0, x_1=1, x_2=2, x_3=3$ 为节点的三次样条函数 $s(x)$，使其满足条件

$$s(x_i)=f(x_i)\quad (i=0,1,2,3)$$
$$s''(x_0)=f''(x_0),\quad s''(x_3)=f''(x_3)$$

26. 给定数表

x	1	2	3	4
$f(x)$	8	6	5	7
$f'(x)$	-1			2

用三弯矩法求以 $x_0=1, x=2, x_2=3, x_3=4$ 为节点的三次样条函数 $s(x)$，使其满足条件：

$$s(x_i)=f(x_i)\quad (i=0,1,2,3)$$
$$s'(x_0)=f'(x_0),\quad s'(x_3)=f'(x_3)$$

27. 设有 $n=4$ 的实序列

$$f_0=1,\quad f_1=1,\quad f_2=3,\quad f_3=1$$

试用 FFT 算法求 $\{f_l\}$ 的离散频谱。

28. 设有周期为 2π 的函数

$$f(x)=\begin{cases}1, & 0\leqslant x\leqslant \pi\\ \dfrac{x}{\pi}, & \pi<x<2\pi\\ 1, & x=2\pi\end{cases}$$

取 $n=8$，试用 FFT 算法求 $\{f_l(l=0,1,\cdots,7)\}$ 的离散频谱。

29. 若 $\{\varphi_j(x)(j=0,1,\cdots)\}$ 是区间 $[a,b]$ 上带权 $\rho(x)$ 的正交函数系，证明：此函数系在区间 $[a,b]$ 上线性无关。

30. 设 $\{g_k(x)\}$ 和 $\{\varphi_k(x)\}$ 都是区间 $[a,b]$ 上带权 $\rho(x)$ 的正交多项式系，试证：必存在非零实数 $c_k(k=0,1,\cdots)$，使得下列关系成立：

$$g_k(x)\equiv c_k\varphi_k(x),\quad x\in[a,b]\quad (k=0,1,\cdots)$$

31. 设 $\{g_k(x)\}$ 是最高次项系数为 1 的区间 $[a,b]$ 上带权 $\rho(x)$ 的正交多项式系。试由正交多项式系的递推关系(5.73)推出如下的递推关系($k\geqslant 1$)：

$$g_{k+1}(x)=(x-\beta_k)g_k(x)-\mu_{k-1}g_{k-1}(x)$$

其中

$$\beta_k=\frac{(xg_k,g_k)}{(g_k,g_k)},\quad \mu_{k-1}=\frac{(g_k,g_k)}{(g_{k-1},g_{k-1})}$$

32. 试构造区间 $[0,1]$ 上带权 $\rho(x)=x$ 的正交多项式组 $\{\varphi_k(x)(k=0,1,2)\}$。

33. 设 $T_k(x)$ 是 k 次 Chebyshev 多项式，证明：$T_{2n}(x)=T_n(2x^2-1)$。

34. 定义内积

$$(f,g)=\int_{-1}^{1}f(x)g(x)\mathrm{d}x$$

在 $H=\mathrm{Span}\{1,x^2,x^4\}$ 中求对于 $f(x)=|x|$ 的最佳平方逼近元素。

35. 求函数 $f(x)=\sin\pi x$ 在区间 $[-1,1]$ 上的三次最佳平方逼近多项式。

36. 求函数 $f(x)=x^3$ 在区间 $[0,1]$ 上的二次最佳平方逼近多项式。

37. 求 a,b,c 的值，使

$$\int_0^{\pi}(\sin x-a-bx-cx^2)^2\mathrm{d}x$$

达到最小。

38. 设 $p(x)=\sum\limits_{j=0}^{n}a_j\varphi_j(x)$ 是在空间 $H=\mathrm{Span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$ 中对于 $f(x)\in C[a,b]$ 的最佳平方逼近元素，试证：

$$(f-p,f-p)=(f,f)-\sum_{j=0}^{n}a_j(\varphi_j,f)$$

39. 求函数 $f(x)=\arccos x$ 按 Chebyshev 多项式展开的 $n=8$ 的部分和。

40. 给定区间 $[0,1]$ 的分划

$$\pi:\quad x_0=0,\quad x_1=\frac{1}{3},\quad x_2=\frac{2}{3},\quad x_3=1$$

在对应于分划 π 的一次样条函数空间 $\mathscr{D}_{1,\pi}$ 中寻求对于 $f(x)=\sqrt{x}$ 的最佳平方逼近元素 $s(x)$。

41. 利用最小二乘原则求一个形如 $y=a+bx^2$ 的经验公式,使它与下列数据拟合

x	19	25	31	38	44
y	19.0	32.3	49.0	73.3	97.8

42. 给定数表

x	−0.75	−0.5	−0.25	0	0.25	0.5	0.75
y	0.33	0.88	1.44	2.00	2.56	3.13	3.71

试分别用一次、二次、三次多项式根据最小二乘原则拟合这些数据,并比较优劣。

43. 欲求一个形如 $s=ct^\lambda$ 的经验公式,使它与实验数据

t	1	2	4	8	16	32	64
s	4.22	4.02	3.85	3.59	3.44	3.02	2.59

相拟合,试用最小二乘法确定其中的参数 c 和 λ。

44. 给定数表

i	1	2	3	4	5	6	7	8	9
x_i	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$	$\frac{3}{4}\pi$	π	$\frac{5}{4}\pi$	$\frac{3}{2}\pi$	$\frac{7}{4}\pi$	2π
y_i	0	1.5	3	1.5	0	−1.5	−3	−1.5	0

试确定 $y(x)=c_1\sin x+c_2\sin 2x+c_3\sin 3x$ 的系数 c_1,c_2 和 c_3,使得

$$\sum_{i=1}^{9}[y(x_i)-y_i]^2=\min$$

45. 试构造在点集{1,2,…,10}上正交的多项式系$\{\varphi_k(x)(k=0,1,2)\}$,其中 $\varphi_k(x)$是 k 次多项式。

46. 试取 $\varphi_r(x)=x^r(r=0,1,2)$,$\psi_0(y)\equiv1$,$\psi_1(y)=y^2$,对第 16 题所给数据(见表 5-6)作乘积型最小二乘拟合,并求拟合精度。

第 6 章　数值积分

为了计算定积分 $\int_a^b f(x)\mathrm{d}x$，只要求出被积函数 $f(x)$ 的一个原函数 $F(x)$，再利用 Newton - Leibniz（牛顿-莱布尼茨）公式

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a)$$

计算，问题就解决了。这是通过解析方法计算定积分。但是，在实际工作中，有很多被积函数是很难求出它的原函数的，甚至不能求出有限形式的原函数。例如，积分

$$\int_a^b \frac{1}{\ln x}\mathrm{d}x,\quad \int_a^b \frac{\sin x}{x}\,\mathrm{d}x$$

等都不能通过解析方法计算。此外，实际问题中的被积函数 $f(x)$ 往往没有解析表达式，只用数表的形式 $(x_i, f(x_i))$ $(i=0,1,\cdots,n)$ 给出 $f(x)$。这种情况更无法采用解析方法计算 $f(x)$ 的定积分。本章将要讨论计算定积分的数值方法——数值积分法。

6.1　求积公式及其代数精度

数值求积公式的一般形式为

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{k=0}^{n} \lambda_k f(x_k) \tag{6.1}$$

式中的 $x_k(k=0,1,\cdots,n)$ 称为求积节点，并且有

$$a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$$

$\lambda_k(k=0,1,\cdots,n)$ 称为求积系数，且与被积函数 $f(x)$ 无关。数值求积公式(6.1)通常简称为求积公式。当求积节点和求积系数确定之后，一个求积公式就被确定了。表达式

$$R_n = \int_a^b f(x)\mathrm{d}x - \sum_{k=0}^{n} \lambda_k f(x_k)$$

被称为求积公式(6.1)的截断误差或余项。

数值求积公式是一种近似方法，因此，要求它对尽可能多的被积函数 f 能准确计算积分 $\int_a^b f(x)\mathrm{d}x$ 的值，这就引出了代数精确度的概念。

定义　对于求积公式(6.1)，当 $f(x)$ 为任何次数不高于 m 的多项式时都成为等式，而当 $f(x)$ 为某个 $m+1$ 次多项式时不能成为等式，则称它具有 m 次代数精度。

容易看出，只要当 $f(x)$ 分别为 $1, x, x^2, \cdots, x^m$ 时，求积公式(6.1)都成为等式，则当 $f(x)$ 为任何次数不高于 m 的多项式时，求积公式(6.1)就必成为等式。这个事实说明了如何判明一个求积公式的代数精度。

例 1　判明以下两个求积公式的代数精度

(1) $\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{1}{2}[f(-1) + 2f(0) + f(1)]$；

(2) $\int_{-1}^{1} f(x)\mathrm{d}x \approx f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right)$。

解 (1) 当 $f(x)$分别为 1 和 x 时，求积公式都成为等式；而当 $f(x)=x^2$ 时，求积公式的左端值不等于右端值，故此求积公式具有一次代数精度。

(2) 当 $f(x)$分别为 $1,x,x^2,x^3$ 时，求积公式都成为等式；而当 $f(x)=x^4$ 时，求积公式的左端值不等于右端值，故此求积公式具有三次代数精度。

代数精度这个概念只是定性地描述一个求积公式的精度高低，并不能定量地表示求积公式的误差大小。

6.2 插值型求积公式

推导求积公式的基本方法是利用插值法。

在区间$[a,b]$内给定求积节点 $x_k(k=0,1,\cdots,n)$，并且

$$a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$$

以所给求积节点为插值节点，构造函数 $f(x)$的 Lagrange 插值多项式

$$p_n(x) = \sum_{k=0}^{n} l_k(x)f(x_k)$$

其中 $l_k(x)$是 Lagrange 插值基函数

$$l_k(x) = \prod_{\substack{i=0\\ i\neq k}}^{n} \frac{x-x_i}{x_k-x_i} \quad (k=0,1,\cdots,n)$$

用 $p_n(x)$近似代替 $f(x)$，在区间$[a,b]$上作定积分，就得到近似等式

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{k=0}^{n} \lambda_k^{(n)} f(x_k) \tag{6.1$_1$}$$

其中

$$\lambda_k^{(n)} = \int_a^b l_k(x)\mathrm{d}x \quad (k=0,1,\cdots,n) \tag{6.2}$$

当求积系数 $\lambda_k^{(n)}$ 由式(6.2)确定时，相应的求积公式(6.1)$_1$ 就称为插值型求积公式。

设函数 $f(x)$在区间$[a,b]$上足够光滑，根据定理 5.1，对于$a\leqslant x\leqslant b$，有

$$f(x) = \sum_{k=0}^{n} l_k(x)f(x_k) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{j=0}^{n}(x-x_j)$$

其中 $\xi=\xi(x)\in(a,b)$。由此得到插值型求积公式(6.1)$_1$ 的截断误差为

$$R_n = \int_a^b f(x)\mathrm{d}x - \sum_{k=0}^{n}\lambda_k^{(n)}f(x_k) = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\left[\prod_{j=0}^{n}(x-x_j)\right]\mathrm{d}x \tag{6.3}$$

其中 $\xi=\xi(x)\in(a,b)$。

例 2 给定求积节点 $x_0=\frac{1}{4}, x_1=\frac{3}{4}$，试推出计算积分 $\int_0^1 f(x)\mathrm{d}x$ 的插值型求积公式，并写出它的截断误差。

解 由公式(6.2)计算求积系数

$$\lambda_0^{(1)} = \int_0^1 \frac{x-x_1}{x_0-x_1}\mathrm{d}x = -\frac{1}{2}\int_0^1 (4x-3)\mathrm{d}x = \frac{1}{2}$$

$$\lambda_1^{(1)} = \int_0^1 \frac{x-x_0}{x_1-x_0}\mathrm{d}x = \frac{1}{2}\int_0^1 (4x-1)\mathrm{d}x = \frac{1}{2}$$

故求积公式为

$$\int_0^1 f(x)\mathrm{d}x \approx \frac{1}{2}\left[f\left(\frac{1}{4}\right)+f\left(\frac{3}{4}\right)\right]$$

由式(6.3)，此求积公式的截断误差为

$$R_1 = \int_0^1 \frac{1}{2}f''(\xi)\left(x-\frac{1}{4}\right)\left(x-\frac{3}{4}\right)\mathrm{d}x$$

其中 $\xi=\xi(x)\in(0,1)$。

定理 6.1　$n+1$ 个节点的插值型求积公式$(6.1)_1$、(6.2)至少具有 n 次代数精度。

证　当 $f(x)$ 为任何次数不高于 n 的多项式时，$f^{(n+1)}(x)\equiv 0$，根据插值型求积公式的截断误差表达式(6.3)可知，此时$R_n=0$，因而等式

$$\int_a^b f(x)\mathrm{d}x = \sum_{k=0}^{n}\lambda_k^{(n)} f(x_k)$$

成立，其中 $\lambda_k^{(n)}$ 由式(6.2)确定。根据代数精度的定义，可知定理的结论成立。

证毕。

推论　对于 $n+1$ 个节点的插值型求积公式的求积系数 $\lambda_k^{(n)}$ $(k=0,1,\cdots,n)$，必满足

$$\sum_{k=0}^{n}\lambda_k^{(n)} = b-a$$

其中 a 和 b 分别是积分下限和上限。

定理 6.2　$n+1$ 个节点的求积公式(6.1)如果具有 n 次或大于 n 次的代数精度，则它是插值型求积公式。

此定理请读者自己证明。

6.3　Newton－Cotes 求积公式

如果节点等距，且 $x_0=a, x_n=b$，即

$$x_k = a+kh \quad (k=0,1,\cdots,n) \quad h=\frac{b-a}{n}$$

则相应的插值型求积公式$(6.1)_1$、(6.2)称为 Newton－Cotes（牛顿-科茨）求积公式，相应的求积系数 $\lambda_k^{(n)}$ 称为 Newton－Cotes 求积系数。

令 $x=a+th$，由式(6.2)得

$$\lambda_k^{(n)} = \int_a^b l_k(x)\mathrm{d}x = \int_a^b\left(\prod_{\substack{j=0\\j\neq k}}^{n}\frac{x-x_j}{x_k-x_j}\right)\mathrm{d}x = \int_0^n\left(\prod_{\substack{j=0\\j\neq k}}^{n}\frac{t-j}{k-j}\right)h\,\mathrm{d}t =$$

$$\frac{(-1)^{n-k}h}{k!(n-k)!}\int_0^n\left[\prod_{\substack{j=0\\j\neq k}}^{n}(t-j)\right]\mathrm{d}t = (b-a)c_k^{(n)} \quad (k=0,1,\cdots,n) \tag{6.4}$$

其中

$$c_k^{(n)} = \frac{(-1)^{n-k}}{k!(n-k!)n}\int_0^n\left[\prod_{\substack{j=0\\j\neq k}}^{n}(t-j)\right]\mathrm{d}t \quad (k=0,1,\cdots,n) \tag{6.5}$$

$c_k^{(n)}$ 称为 Cotes 系数，它只与 k 和 n 有关，与被积函数 $f(x)$ 以及积分区间 $[a,b]$ 都无关。根据式(6.5)已算出 Cotes 系数表($n\leqslant 8$)，见表 6-1。

表 6-1 Cotes 系数表

n	$c_k^{(n)}$								
1	$\frac{1}{2}$	$\frac{1}{2}$							
2	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$						
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$					
4	$\frac{7}{90}$	$\frac{16}{45}$	$\frac{2}{15}$	$\frac{16}{45}$	$\frac{7}{90}$				
5	$\frac{19}{288}$	$\frac{25}{96}$	$\frac{25}{144}$	$\frac{25}{144}$	$\frac{25}{96}$	$\frac{19}{288}$			
6	$\frac{41}{840}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{34}{105}$	$\frac{9}{280}$	$\frac{9}{35}$	$\frac{41}{840}$		
7	$\frac{751}{172\,80}$	$\frac{357\,7}{172\,80}$	$\frac{132\,3}{172\,80}$	$\frac{298\,9}{172\,80}$	$\frac{298\,9}{172\,80}$	$\frac{132\,3}{172\,80}$	$\frac{357\,7}{172\,80}$	$\frac{751}{172\,80}$	
8	$\frac{989}{283\,50}$	$\frac{588\,8}{283\,50}$	$\frac{-928}{283\,50}$	$\frac{104\,96}{283\,50}$	$\frac{-454\,40}{283\,50}$	$\frac{104\,96}{283\,50}$	$\frac{-928}{283\,50}$	$\frac{588\,8}{283\,50}$	$\frac{989}{283\,50}$

当 n 给定之后，由表 6-1 和式(6.4)即可得到 Newton-Cotes 求积系数 $\lambda_k^{(n)}(k=0,1,\cdots,n)$，从而得到 Newton-Cotes 求积公式

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{k=0}^{n}\lambda_k^{(n)} f\left(a+k\frac{b-a}{n}\right) \tag{6.6}$$

由式(6.3)且令 $x=a+th$，可得到求积公式(6.6)的截断误差

$$R_n = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\left[\prod_{j=0}^{n}(x-x_j)\right]\mathrm{d}x =$$

$$\frac{h^{n+2}}{(n+1)!}\int_0^n f^{(n+1)}(\xi)\left[\prod_{j=0}^{n}(t-j)\right]\mathrm{d}t \tag{6.7}$$

其中 $\xi=\xi(a+th)\in(a,b)$。

定理 6.3 当 n 为偶数时，$n+1$ 个节点的 Newton-Cotes 求积公式的代数精度至少是 $n+1$。

证 由定理 6.1 可知，$n+1$ 个节点的 Newton-Cotes 求积公式至少具有 n 次代数精度。下面只须证明，当 n 为偶数并且 $f(x)=x^{n-1}$ 时，式(6.7)所表示的截断误差 $R_n=0$。

由 $f(x)=x^{n+1}$ 得 $f^{(n-1)}(x)=(n+1)!$，因而

$$R_n = h^{n+2}\int_0^n t(t-1)\cdots(t-n)\mathrm{d}t$$

令 $t=u+\frac{n}{2}$，因 n 是偶数，故 $\frac{n}{2}$ 是整数，于是有

$$R_n = h^{n+2}\int_{-\frac{n}{2}}^{\frac{n}{2}}\left(u+\frac{n}{2}\right)\left(u+\frac{n}{2}-1\right)\cdots(u+1)u(u-1)\cdots$$

$$\left(u-\frac{n}{2}+1\right)\left(u-\frac{n}{2}\right)\mathrm{d}u =$$

$$h^{n+2}\int_{-\frac{n}{2}}^{\frac{n}{2}} u(u^2-1)(u^2-4)\cdots\left(u^2-\frac{n^2}{4}\right)\mathrm{d}u = 0$$

证毕。

下面根据表 6-1、式(6.4)及式(6.6)列出几个常用的Newton-Cotes 求积公式。

1. 梯形公式

$n=1$ 的 Newton-Cotes 求积公式

$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{2}[f(a)+f(b)] \tag{6.8}$$

称为梯形公式，其几何意义是明显的。

设 $f(x)$在区间$[a,b]$上有二阶连续导数。由式(6.7)且$n=1,h=b-a$，得到梯形公式(6.8)的截断误差

$$R_1 = \frac{(b-a)^3}{2}\int_0^1 f''(\xi)t(t-1)\mathrm{d}t$$

其中 $\xi=\xi(a+th)\in(a,b)$。设 $\xi(a+th)$在 $0\leqslant t\leqslant 1$ 上连续。由于 $f''(\xi(a+th))$在 $0\leqslant t\leqslant 1$ 上连续以及 $t(t-1)$在区间 $0<t<1$ 内不变号，故根据积分中值定理，必存在 $\tilde{t}\in[0,1]$，使得下式成立：

$$\int_0^1 f''(\xi)t(t-1)\mathrm{d}t = f''(\xi(a+\tilde{t}h))\int_0^1 t(t-1)\mathrm{d}t = -\frac{f''(\eta)}{6}$$

其中 $\eta=\xi(a+\tilde{t}h)\in(a,b)$。因此，梯形公式(6.8)的截断误差为

$$R_1 = -\frac{(b-a)^3}{12}f''(\eta),\quad \eta\in(a,b) \tag{6.9}$$

从式(6.9)看出，梯形公式(6.8)具有一次代数精度。

2. Simpson 公式

$n=2$ 的 Newton-Cotes 求积公式

$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right] \tag{6.10}$$

称为 Simpson(辛普生)公式，又叫抛物线公式，其几何意义也是明显的。

如果 $f(x)$在区间$[a,b]$上有三阶连续导数，则由式(6.7)($n=2$)可表示 Simpson 公式(6.10)的截断误差 R_2。但是，还可推出更便于使用的 R_2 表达式。

设 $f(x)$在区间$[a,b]$上有四阶连续导数。根据定理 6.3，Simpson 公式(6.10)至少具有三次代数精度。因此，只要把 $f(x)$表示为某个三次插值多项式与其插值公式余项之和，然后再作积分便可推出 Simpson 公式(6.10)的截断误差。令取插值条件

$$p_3(a) = f(a),\quad p_3\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right)$$

$$p_3(b) = f(b),\quad p'_3\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right)$$

可构造出 $f(x)$的三次 Hermite 插值多项式 $p_3(x)$，并且由定理5.3可知

$$f(x) = p_3(x) + \frac{f^{(4)}(\xi)}{4!}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)$$

当 $x\in[a,b]$时，$\xi=\xi(x)\in(a,b)$。由上式，得

$$\int_a^b f(x)\mathrm{d}x-\int_a^b p_3(x)\mathrm{d}x=\int_a^b \frac{f^{(4)}(\xi)}{4!}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x$$

利用 Simpson 公式(6.10)可知

$$\int_a^b p_3(x)\mathrm{d}x=\frac{b-a}{6}\left[p_3(a)+4p_3\left(\frac{a+b}{2}\right)+p_3(b)\right]=$$

$$\frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$$

因而有

$$R_2=\int_a^b f(x)\mathrm{d}x-\frac{b-a}{2}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]=$$

$$\int_a^b \frac{f^{(4)}(\xi)}{4!}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x$$

其中 $\xi=\xi(x)\in(a,b)$。设 $\xi(x)\in C[a,b]$，则由于 $f^{(4)}(\xi(x))$ 在 $a\leqslant x\leqslant b$ 上连续，$(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)$在$(a,b)$内不变号，故由积分中值定理得知，存在 $\tilde{\eta}\in[a,b]$使得下式成立：

$$R_2=\frac{f^{(4)}(\xi(\tilde{\eta}))}{4!}\int_a^b(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x=-\frac{(b-a)^5}{2\ 880}f^{(4)}(\eta) \qquad (6.11)$$

其中 $\eta=\xi(\tilde{\eta})\in(a,b)$。

从式(6.11)看出，Simpson 公式(6.10)具有三次代数精度。

3. Simpson 3/8 公式

$n=3$ 的 Newton - Cotes 求积公式

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{8}\left[f(a)+3f\left(\frac{2a+b}{3}\right)+3f\left(\frac{a+2b}{3}\right)+f(b)\right]$$

称为 Simpson 3/8 公式。如果 $f^{(4)}(x)$在$[a,b]$上连续，则此求积公式的截断误差为

$$R_3=-\frac{(b-a)^5}{6\ 480}f^{(4)}(\eta),\quad \eta\in(a,b)$$

4. Cotes 公式

$n=4$ 的 Newton - Cotes 求积公式

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{90}\left[7f(a)+32f\left(\frac{3a+b}{4}\right)+12f\left(\frac{a+b}{2}\right)+32f\left(\frac{a+3b}{4}\right)+7f(b)\right]$$

称为 Cotes 求积公式。如果 $f^{(6)}(x)$在$[a,b]$上连续，则此求积公式的截断误差为

$$R_4=-\frac{(b-a)^7}{1\ 935\ 360}f^{(6)}(\eta),\quad \eta\in(a,b)$$

例 3 试分别使用梯形公式和 Simpson 公式计算积分 $\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x$ 的近似值，并估计截断误差。

解 用梯形公式计算，得

$$\int_1^2 e^{\frac{1}{x}}dx \approx \frac{2-1}{2}(e+e^{\frac{1}{2}}) = 2.1835$$

$$f(x) = e^{\frac{1}{x}},\quad f'(x) = -\frac{1}{x^2}e^{\frac{1}{x}}$$

$$f''(x) = \left(\frac{2}{x^3}+\frac{1}{x^4}\right)e^{\frac{1}{x}}$$

$$\max_{1\leqslant x\leqslant 2} |f''(x)| = f''(1) = 8.1548$$

截断误差估计为

$$|R_1| \leqslant \frac{(2-1)^3}{12}\max_{1\leqslant x\leqslant 2}|f''(x)| = 0.6796$$

用 Simpson 公式计算，得

$$\int_1^2 e^{\frac{1}{x}}dx \approx \frac{2-1}{6}(e+4e^{\frac{1}{1.5}}+e^{\frac{1}{2}}) = 2.0263$$

$$f^{(4)}(x) = \left(\frac{1}{x^8}+\frac{12}{x^7}+\frac{36}{x^6}+\frac{24}{x^5}\right)e^{\frac{1}{x}}$$

$$\max_{1\leqslant x\leqslant 2}|f^{(4)}(x)| = f^{(4)}(1) = 198.43$$

截断误差估计为

$$|R_2| \leqslant \frac{(2-1)^5}{2\,880}\max_{1\leqslant x\leqslant 2}|f^{(4)}(x)| = 0.06890$$

6.4　Newton - Cotes 求积公式的收敛性与数值稳定性

记

$$I(f) = \int_a^b f(x)dx,\quad I_n(f) = \sum_{k=0}^n \lambda_k^{(n)} f\left(a+k\frac{b-a}{n}\right)$$

其中 $\lambda_k^{(n)}$ $(k=0,1,\cdots,n)$是 Newton - Cotes 求积系数。今考察是否对任何在$[a,b]$上可积的函数 $f(x)$都有

$$\lim_{n\to\infty} I_n(f) = I(f)$$

这是求积公式(6.6)的收敛性问题。

先看一个例子，$f(x)=\dfrac{1}{1+x^2}$，$[a,b]=[-4,4]$，此时有

$$I(f) = 2\arctan 4 \approx 2.6516$$

$I_n(f)$的一些计算结果见表 6 - 2。

表 6 - 2　$I_n(f)$数值表

n	$I_n(f)$
2	5.490 2
4	2.277 6
6	3.328 8
8	1.941 1
10	3.595 6

从表 6 - 2 看出，当 $n\to\infty$时，$I_n(f)$不收敛于 $I(f)$。这个例子说明，Newton - Cotes 求积公式并不是对所有在区间$[a,b]$上可积的函数都收敛。

设计算 $f(x_k)$时有舍入误差 $\varepsilon_k = f(x_k)-\tilde{f}(x_k)$ $(k=0,1,\cdots,n)$，并设对任何 k 有$|\varepsilon_k|\leqslant\varepsilon$

(常数),则计算 $I_n(f)$时,由 ε_k 引起的积累误差为

$$\eta_n = I_n(f) - I_n(\tilde{f}) = \sum_{k=0}^{n} \lambda_k^{(n)} \varepsilon_k$$

今考察当 $n \to \infty$时,$|\eta_n|$是否有界,这是求积公式(6.6)的数值稳定性问题。由

$$|\eta_n| \leqslant \sum_{k=0}^{n} |\lambda_k^{(n)}| |\varepsilon_k| \leqslant \varepsilon \sum_{k=0}^{n} |\lambda_k^{(n)}|$$

可知,如果对任何 n,$\sum_{k=0}^{n} |\lambda_k^{(n)}| \leqslant K$(常数),则求积公式(6.6)就具有数值稳定性。但是,理论上已经证明,对于 Newton - Cotes 求积系数 $\lambda_k^{(n)}$,当 $n \to \infty$时,$\sum_{k=0}^{n} |\lambda_k^{(n)}|$ 是无界的(证明见文献[3]第 192 页)。因此,Newton - Cotes 求积公式(6.6)的数值稳定性是没有保证的。

综上所述,多节点的 Newton - Cotes 求积公式不宜使用,用得较多的是 $n=1,2,4$ 的情形。

6.5 复化求积法

前面已经论述过,不宜使用多节点的 Newton - Cotes 求积公式。但是,当积分区间的长度较大时,少节点的 Newton - Cotes 求积公式的截断误差比较大。为了提高计算积分的精确度,可以把积分区间分为若干个子区间。在每个子区间上的积分使用少节点的 Newton - Cotes 求积公式计算,然后把结果相加,这就是复化求积法,所得的求积公式称为复化求积公式。

6.5.1 复化梯形公式与复化 Simpson 公式

设 $f(x)$在区间$[a,b]$上有二阶连续导数,取等距节点

$$x_k = a + kh \quad (k = 0,1,\cdots,n) \quad h = \frac{b-a}{n}$$

在每个子区间$[x_k,x_{k+1}]$上的积分使用梯形公式(6.8)及其截断误差式(6.9),得

$$\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x = \frac{h}{2}[f(x_k) + f(x_{k+1})] - \frac{h^3}{12} f''(\eta_k)$$

其中 $\eta_k \in (x_k, x_{k-1})$。于是有

$$\int_a^b f(x)\mathrm{d}x = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x = \frac{h}{2} \sum_{k=0}^{n-1} [f(x_k) + f(x_{k+1})] - \frac{h^3}{12} \sum_{k=0}^{n-1} f''(\eta_k)$$

略去上式右端第二个和式,并经整理得到

$$\int_a^b f(x)\mathrm{d}x \approx \frac{h}{2}\Big[f(a) + f(b) + 2\sum_{k=1}^{n-1} f(a+kh)\Big] \tag{6.12}$$

称式(6.12)为复化梯形公式,它的截断误差为

$$R_T = -\frac{h^3}{12} \sum_{k=0}^{n-1} f''(\eta_k), \quad \eta_k \in (x_k, x_{k-1})$$

因 $f''(x)$在$[a,b]$上连续,故存在 $\eta \in [a,b]$,使

$$f''(\eta) = \frac{1}{n} \sum_{k=0}^{n-1} f''(\eta_k)$$

因此复化梯形公式(6.12)的截断误差可表示为

$$R_{\mathrm{T}} = -\frac{(b-a)^3}{12n^2}f''(\eta) = -\frac{b-a}{12}h^2 f''(\eta) \tag{6.13}$$

其中 $\eta\in[a,b]$。

如果 $f(x)$在$[a,b]$上有 $2r+2$ 阶连续导数，则复化梯形公式(6.12)的截断误差还可表示为

$$\begin{aligned}R_{\mathrm{T}} = {} & \frac{B_2}{2!}h^2[f'(a)-f'(b)] + \frac{B_4}{4!}h^4[f'''(a)-f'''(b)] + \cdots + \\ & \frac{B_{2r}}{(2r)!}h^{2r}[f^{(2r-1)}(a)-f^{(2r-1)}(b)] - \\ & \frac{B_{2r+2}(b-a)}{(2r+2)!}h^{(2r+2)}f^{(2r-2)}(\eta)\end{aligned} \tag{6.13$_1$}$$

其中 $\eta\in[a,b]$，式中的 B_j 是 Bernoulli(伯努利)数，其前几个数是

$$B_0 = 1,\quad B_1 = -\frac{1}{2},\quad B_2 = \frac{1}{6},\quad B_4 = -\frac{1}{30}$$

$$B_6 = \frac{1}{42},\quad B_8 = -\frac{1}{30},\quad B_{10} = \frac{5}{66},\quad B_{12} = -\frac{691}{2\,730}$$

对一切 $j\geqslant 3$ 的奇数 $B_j=0$。公式$(6.13)_1$ 的证明参见文献[3]第 213～215 页。

利用定积分定义容易证明，只要函数 $f(x)$在区间$[a,b]$上可积，则当 $n\to\infty$时，复化梯形公式(6.12)右端(称为复化梯形值)收敛于积分值 $\int_a^b f(x)\mathrm{d}x$ 。

此外，求积公式(6.12)的求积系数 $\lambda_k(k=0,1,\cdots,n)$总满足

$$\sum_{k=0}^{n}|\lambda_k| = nh = b-a$$

因而公式(6.12)具有数值稳定性。

设 $f(x)$在区间$[a,b]$上有四阶连续导数。取 $2m+1$ 个等距节点

$$x_k = a+kh\quad(k=0,1,\cdots,2m)\quad h=\frac{b-a}{2m}$$

在子区间$[x_{2i-2},x_{2i}](i=1,2,\cdots,m)$上的积分使用 Simpson 公式(6.10)及其截断误差(6.11)，得

$$\begin{aligned}\int_{x_{2i-2}}^{x_{2i}} f(x)\mathrm{d}x = {} & \frac{h}{3}[f(x_{2i-2})+4f(x_{2i-1})+f(x_{2i})] - \\ & \frac{(2h)^5}{2\,880}f^{(4)}(\eta_i),\quad \eta_i\in(x_{2i-2},x_{2i})\end{aligned}$$

于是有

$$\int_a^b f(x)\mathrm{d}x = \sum_{i=1}^{m}\frac{h}{3}[f(x_{2i-2})+4f(x_{2i-1})+f(x_{2i})] - \frac{(2h)^5}{2\,880}\sum_{i=1}^{m}f^{(4)}(\eta_i)$$

略去上式右端第二个和式，并经整理得到

$$\int_a^b f(x)\mathrm{d}x \approx \frac{h}{3}\left[f(a)+f(b)+4\sum_{i=1}^{m}f(x_{2i-1})+2\sum_{i=1}^{m-1}f(x_{2i})\right] \tag{6.14}$$

称式(6.14)为复化 Simpson 公式。因 $f^{(4)}(x)$ 在$[a,b]$ 上连续，故存在 $\eta\in[a,b]$，使

$$f^{(4)}(\eta) = \frac{1}{m}\sum_{i=1}^{m} f^{(4)}(\eta_i)$$

因此，复化 Simpson 公式(6.14)的截断误差为

$$R_S = -\frac{(2h)^5}{2\ 880} m f^{(4)}(\eta) = -\frac{(b-a)^5}{2\ 880 m^4} f^{(4)}(\eta) =$$

$$-\frac{b-a}{180} h^4 f^{(4)}(\eta), \quad \eta \in [a,b] \tag{6.15}$$

利用定积分定义也容易证明，只要函数 $f(x)$ 在区间$[a,b]$上可积，则当 $m \to \infty$ 时，复化 Simpson 公式(6.14)右端(称为复化 Simpson 值)收敛于积分值 $\int_a^b f(x)\mathrm{d}x$ 。由于求积公式(6.14)的求积系数 $\lambda_k (k = 0,1,\cdots,2m)$ 总满足

$$\sum_{k=0}^{n} |\lambda_k| = 2mh = b - a$$

因而公式(6.14)具有数值稳定性。

易知，复化梯形公式(6.12)和复化 Simpson 公式(6.14)都不属于插值型求积公式。

例 4 用 11 个节点的复化 Simpson 公式计算积分 $\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x$ 的近似值，并估计截断误差。

解 $m=5, h=0.1$，求积节点为

$$x_k = 1 + 0.1k \quad (k = 0,1,\cdots,10)$$

由公式(6.14)，得

$$\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x \approx \frac{0.1}{3}\left(\mathrm{e} + \mathrm{e}^{\frac{1}{2}} + 4\sum_{i=1}^{5} \mathrm{e}^{\frac{1}{x_{2i-1}}} + 2\sum_{i=1}^{4} \mathrm{e}^{\frac{1}{x_{2i}}}\right) = 2.020\ 077$$

由式(6.15)得截断误差估计

$$|R_S| \leqslant \frac{2-1}{180}(0.1)^4 \max_{1 \leqslant x \leqslant 2} |f^{(4)}(x)| = \frac{(0.1)^4}{180} \times 198.43 = 0.000\ 11$$

由此可知，2.020 07 作为积分 $\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x$ 的近似值能有四位有效数字。

例 5 如果用复化梯形公式计算积分 $\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x$ 的近似值 I_n，并要求 I_n 至少具有四位有效数字，则须用多少个节点的复化梯形公式？(不计舍入误差)

解 由

$$\sqrt{\mathrm{e}} < \int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x < \mathrm{e}$$

可知该积分值的第一位非零数字在个位，又因要求近似值 I_n 至少具有四位有效数字，所以 I_n 的截断误差 R_T 应满足

$$|R_T| \leqslant 0.000\ 5$$

由式(6.13)可知，只须

$$\frac{(2-1)^3}{12n^2} \max_{1 \leqslant x \leqslant 2} |f''(x)| = \frac{8.154\ 8}{12n^2} \leqslant 0.000\ 5$$

$$n \geqslant 37$$

所以，至少须用 38 个节点的复化梯形公式计算。

从以上两个例子可看出，为达到相同的精度水平，使用复化梯形公式所需的计算量比使用

复化 Simpson 公式的计算量大。

还可以推出其他复化 Newton－Cotes 求积公式。常用的还有复化 Cotes 公式。

定义　设复化求积公式为

$$I(f)=\int_a^b f(x)\mathrm{d}x \approx S_n(f)$$

其中 n 为区间$[a,b]$的等分数，且对任何$[a,b]$上可积的函数 $f(x)$有$\lim\limits_{n\to\infty} S_n(f)=I(f)$。记 $h=\dfrac{b-a}{n}$，若存在常数 $p\geqslant 1$ 和 $c>0$，使对任何 n 有

$$|I(f)-S_n(f)|\leqslant ch^p$$

成立，则称积分近似值序列$\{S_n(f)\}$是 p 阶收敛的。

由式(6.13)和式(6.15)可知，若 $f(x)$在区间$[a,b]$上二次连续可微，则它的复化梯形值序列是二阶收敛的；若 $f(x)$在区间$[a,b]$上四次连续可微，则它的复化 Simpson 值序列是四阶收敛的。

6.5.2　区间逐次分半法

使用复化求积公式计算积分近似值，节点数目越多，截断误差越小。但是，节点多计算量就大。如果想要利用截断误差的表达式预先确定满足精度要求的最少节点数，那么，困难在于寻找被积函数的高阶导数在区间$[a,b]$上的界，解决上述矛盾的一个有效方法是让节点数目从少到多地变化，即让步长可变。

区间逐次分半法就是根据规定的精度要求，在计算过程中把积分区间逐次分半，并利用前后两次计算结果之差来判别误差的大小，从而得到满足精度要求的积分近似值。下面针对复化梯形公式讨论区间逐次分半法。

用 T_m 表示积分区间$[a,b]$被分为 $n=2^m$ 等分后所形成的复化梯形值，这时步长 $h_m=(b-a)/2^m$。由复化梯形公式(6.12)的右端，得

$$T_{m-1}=\frac{h_{m-1}}{2}\left[f(a)+f(b)+2\sum_{k=1}^{2^{m-1}-1} f(a+kh_{m-1})\right]$$

$$\begin{aligned}T_m=&\frac{h_m}{2}\left[f(a)+f(b)+2\sum_{k=1}^{2^m-1} f(a+kh_m)\right]=\\&\frac{h_m}{2}\left[f(a)+f(b)+2\sum_{i=1}^{2^{m-1}-1} f(a+2ih_m)+\right.\\&\left.2\sum_{i=1}^{2^{m-1}} f(a+(2i-1)h_m)\right]\end{aligned}$$

利用 $h_{m-1}=2h_m$ 以及 T_{m-1} 的表达式，由上式可得下列递推公式：

$$\begin{cases}T_0=\dfrac{b-a}{2}[f(a)+f(b)]\\[2ex]T_m=\dfrac{1}{2}T_{m-1}+h_m\displaystyle\sum_{i=1}^{2^{m-1}} f(a+(2i-1)h_m)\quad (m=1,2,\cdots)\end{cases}\tag{6.16}$$

由递推公式(6.16)形成的序列$\{T_m\}$称为积分区间逐次分半的复化梯形值序列。根据 6.5.1 小节的讨论可知，只要函数 $f(x)$在区间$[a,b]$上可积，序列$\{T_m\}$就收敛于积分 $I(f)$。

由截断误差表达式(6.13),可知

$$I(f)-T_m=-\frac{(b-a)^3}{12\times 2^{2m}}f''(\eta_1),\quad \eta_1\in[a,b]$$

$$I(f)-T_{m+1}=-\frac{(b-a)^3}{12\times 2^{2m-2}}f''(\eta_2),\quad \eta_2\in[a,b]$$

当 m 较大时,$f''(\eta_1)\approx f''(\eta_2)$,由此可推出

$$I(f)-T_m\approx\frac{4}{3}(T_{m+1}-T_m)$$

因此,可预先给定 T_m 的绝对误差限 $\varepsilon>0$,当满足

$$|T_{m+1}-T_m|<\frac{3}{4}\varepsilon$$

时停止计算,并认为 $T_m\approx I(f)$ 已满足精度要求。

对 Simpson 公式,也可以进行区间逐次分半而得到复化 Simpson 值序列。

6.6 Romberg 积分法

6.6.1 Richardson 外推技术

设有一个常数 F^*,由一个依赖于 $h>0$ 的算法 $F(h)$ 去逼近它,其中 F^* 与 h 无关,并已知 $F(h)$ 逼近 F^* 的截断误差为

$$F^*-F(h)=a_1h^{p_1}+a_2h^{p_2}+\cdots+a_kh^{p_k}+\cdots \tag{6.17}$$

其中 $a_k(k=1,2,\cdots)$ 是与 h 无关的常数,且

$$0<p_1<p_2<\cdots<p_{k-1}<p_k<\cdots$$

式(6.17)说明,$F(h)$ 逼近 F^* 的精度是 h^{p_1} 级的,或说是 p_1 阶的。现在提出问题,能否利用 $F(h)$ 构造出一个新的算法 $F_1(h)$,使 $F_1(h)$ 逼近 F^* 的精度比 p_1 阶更高,例如 p_2 阶等。下面就讨论此问题。

取一正数 $q,q\neq 1$,根据式(6.17)有

$$F^*-F(qh)=a_1(qh)^{p_1}+a_2(qh)^{p_2}+\cdots+a_k(qh)^{p_k}+\cdots$$

用 q^{p_1} 乘式(6.17)两端,得

$$q^{p_1}[F^*-F(h)]=q^{p_1}(a_1h^{p_1}+a_2h^{p_2}+\cdots+a_kh^{p_k}+\cdots)$$

将上述两式相减,得

$$(1-q^{p_1})F^*-[F(qh)-q^{p_1}F(h)]=$$
$$a_2(q^{p_2}-q^{p_1})h^{p_2}+\cdots+a_k(q^{p_k}-q^{p_1})h^{p_k}+\cdots$$

或得

$$F^*-\frac{F(qh)-q^{p_1}F(h)}{1-q^{p_1}}=a_2\frac{q^{p_2}-q^{p_1}}{1-q^{p_1}}h^{p_2}+\cdots+a_k\frac{q^{p_k}-q^{p_1}}{1-q^{p_1}}h^{p_k}+\cdots=$$
$$a_2^{(1)}h^{p_2}+\cdots+a_k^{(1)}h^{p_k}+\cdots \tag{6.18}$$

其中

$$a_2^{(1)}=a_2\frac{q^{p_2}-q^{p_1}}{1-q^{p_1}},\cdots,a_k^{(1)}=a_k\frac{q^{p_k}-q^{p_1}}{1-q^{p_1}},\cdots$$

都是与 h 无关的常数。令

$$F_1(h)=\frac{F(qh)-q^{p_1}F(h)}{1-q^{p_1}}$$

那么，$F_1(h)$逼近 F^* 的精度已提高到 p_2 阶。

类似地，令

$$F_2(h)=\frac{F_1(qh)-q^{p_2}F_1(h)}{1-q^{p_2}}$$

则 $F_2(h)$逼近 F^* 的精度提高到 p_3 阶。

定义 F_j 为

$$\begin{cases} F_0(h)=F(h) \\ F_j(h)=\dfrac{F_{j-1}(qh)-q^{p_j}F_{j-1}(h)}{1-q^{p_j}} \quad (j=1,2,\cdots) \end{cases} \tag{6.19}$$

则 $F_j(h)$逼近 F^* 的截断误差由下面定理指明。

定理 6.4　若 $F(h)$逼近 F^* 的截断误差由式(6.17)给出，那么，由式(6.19)定义的 $F_j(h)$ 逼近 F^* 的截断误差为

$$F^*-F_j(h)=a_{j+1}^{(j)}h^{p_{j+1}}+a_{j+2}^{(j)}h^{p_{j+2}}+\cdots$$

其中 $a_k^{(j)}(k\geqslant j+1)$是与 h 无关的常数。

证　用归纳法证明。当 $j=1$ 时，由式(6.19)有

$$F_1(h)=\frac{F_0(qh)-q^{p_1}F_0(h)}{1-q^{p_1}}=\frac{F(qh)-q^{p_1}F(h)}{1-q^{p_1}}$$

从关系式(6.18)可知定理结论正确。假定 $j=r-1$ 时定理结论正确，那么有

$$F^*-F_{r-1}(h)=a_r^{(r-1)}h^{p_r}+a_{r+1}^{(r-1)}h^{p_{r+1}}+a_{r+2}^{(r-1)}h^{p_{r+2}}+\cdots \tag{6.20}$$

其中 $a_k^{(r-1)}(k\geqslant r)$是与 h 无关的常数。用 qh 替代上式中的 h，得

$$F^*-F_{r-1}(qh)=a_r^{(r-1)}(qh)^{p_r}+a_{r+1}^{(r-1)}(qh)^{p_{r+1}}+a_{r+2}^{(r-1)}(qh)^{p_{r+2}}+\cdots$$

用 q^{p_r} 乘式(6.20)两边所得的式子与上式相减，并除以$(1-q^{p_r})$就得到

$$F^*-\frac{F_{r-1}(qh)-q^{p_r}F_{r-1}(h)}{1-q^{p_r}}=a_{r+1}^{(r)}h^{p_{r+1}}+a_{r+2}^{(r)}h^{p_{r+2}}+\cdots$$

其中 $a_k^{(r)}(k\geqslant r+1)$是与 h 无关的常数。由于

$$\frac{F_{r-1}(qh)-q^{p_r}F_{r-1}(h)}{1-q^{p_r}}=F_r(h)$$

因而当 $j=r$ 时定理结论也正确。

证毕。

式(6.19)被称为 Richardson(李查逊)外推技术或外推算法。实际上，这种外推技术是由已知的序列

$$F(h),\ F(qh),\ F(q^2h),\ \cdots$$

通过式(6.19)产生第二个序列

$$F_1(h),F_1(qh),F_1(q^2h),\cdots$$

又通过式(6.19)产生第三个序列

$$F_2(h),F_2(qh),F_2(q^2h),\cdots$$

依此类推可产生第四、第五等很多个序列。若序列$\{F(q^mh)\}$收敛于 F^*，并且是 p_1 阶收敛的，

则序列$\{F_1(q^m h)\}$也收敛于F^*，并且是p_2阶收敛的，序列$\{F_2(q^m h)\}$是p_3阶收敛的，依此类推。

Richardson 外推技术的计算顺序可按表 6-3 所示执行，表中的①，②，…表示计算顺序。

表 6-3 Richardson 外推技术的计算顺序

①	$F_0(h)$						
②	$F_0(qh)$	③	$F_1(h)$				
④	$F_0(q^2h)$	⑤	$F_1(qh)$	⑥	$F_2(h)$		
⑦	$F_0(q^3h)$	⑧	$F_1(q^2h)$	⑨	$F_2(qh)$	⑩	$F_3(h)$
	⋮		⋮		⋮		⋮

根据表 6-3 所示的计算顺序，Richardson 外推技术的算法如下：

$$\begin{cases} \text{令 } F_0(h) = F(h) \\ \text{对于 } m = 1,2,\cdots \text{ 计算} \\ (1)\ F_0(q^m h) = F(q^m h) \\ (2)\ F_j(q^{m-j}h) = \dfrac{F_{j-1}(q^{m-j+1}h) - q^{p_j}F_{j-1}(q^{m-j}h)}{1-q^{p_j}} \\ \qquad (j = 1,2,\cdots,m) \end{cases} \tag{6.21}$$

6.6.2 Romberg 积分法

由递推公式(6.16)形成的复化梯形值序列$\{T_m\}$虽然收敛于积分$I(f)$，但它只是二阶收敛的。现在要在序列$\{T_m\}$的基础上通过 Richardson 外推技术产生新的序列，使其以更高阶的收敛速度收敛于积分$I(f)$。为此，设$f(x)$在区间$[a,b]$上足够光滑。

把T_m记为$T_m^{(0)}$，它是步长为$h_m=(b-a)/2^m$的复化梯形值。今取

$$q = \frac{1}{2},\quad h = b-a,\quad F_0(q^m h) = T_m^{(0)}\quad (m = 0,1,\cdots)$$

根据复化梯形公式的截断误差表达式$(6.13)_1$，有

$$I(f) - T_0^{(0)} = a_1h^2 + a_2h^4 + \cdots + a_kh^{2k} + \cdots$$

其中$a_k(k=1,2,\cdots)$是与h无关的常数。

利用 Richardson 外推算法(6.21)，得到如下的求积方法(只产生四个序列$\{T_m^{(0)}\}$，$\{T_m^{(1)}\}$，$\{T_m^{(2)}\}$，$\{T_m^{(3)}\}$)：

$$\begin{cases} T_0^{(0)} = \dfrac{b-a}{2}[f(a)+f(b)] \\ \text{对于 } m = 1,2,\cdots \text{ 计算} \\ (1)\ h_m = \dfrac{b-a}{2^m} \\ (2)\ T_m^{(0)} = \dfrac{1}{2}T_{m-1}^{(0)} + h_m\sum_{i=1}^{2^{m-1}} f(a+(2i-1)h_m) \\ (3)\ T_{m-j}^{(j)} = \dfrac{T_{m-j+1}^{(j-1)} - \left(\frac{1}{2}\right)^{2j}T_{m-j}^{(j-1)}}{1-\left(\frac{1}{2}\right)^{2j}} = \\ \qquad \dfrac{4^jT_{m-j+1}^{(j-1)} - T_{m-j}^{(j-1)}}{4^j-1}\quad [j = 1,2,\cdots,\min(3,m)] \end{cases} \tag{6.22}$$

称方法(6.22)为 Romberg(龙贝格)积分法。它的计算顺序如表 6-4 所示。

表 6-4　Romberg 积分法的计算顺序

① $T_0^{(0)}$			
② $T_1^{(0)}$	③ $T_0^{(1)}$		
④ $T_2^{(0)}$	⑤ $T_1^{(1)}$	⑥ $T_0^{(2)}$	
⑦ $T_3^{(0)}$	⑧ $T_2^{(1)}$	⑨ $T_1^{(2)}$	⑩ $T_0^{(3)}$
⑪ $T_4^{(0)}$	⑫ $T_3^{(1)}$	⑬ $T_2^{(2)}$	⑭ $T_1^{(3)}$
⋮	⋮	⋮	⋮

表 6-4 中的第一列 $\{T_m^{(0)}\}$ 就是复化梯形值序列，是二阶收敛的。由于第一列收敛于积分 $I(f)$，因而其余各列都收敛于积分 $I(f)$。第二列 $\{T_i^{(1)}\}$ 是四阶收敛的，它是复化 Simpson 值序列，这是因为

$$T_m^{(1)} = \frac{4T_{m+1}^{(0)} - T_m^{(0)}}{3}$$

恰好是步长为 $h_{m+1} = (b-a)/2^{m+1}$ 的复化 Simpson 值。第三列 $\{T_m^{(2)}\}$ 是六阶收敛的，称为 Cotes 值序列。第四列 $\{T_m^{(3)}\}$ 是八阶收敛的，称为 Romberg 值序列。

Romberg 积分法是一个迭代过程，控制迭代结束的准则是

$$\frac{|T_m^{(3)} - T_{m-1}^{(3)}|}{|T_m^{(3)}|} \leqslant \varepsilon$$

$\varepsilon > 0$ 是预先给定的精度水平。当满足准则的条件时，结束迭代，当前的 $T_m^{(3)}$ 就是所求的积分 $I(f)$ 的近似值。

例 6　用 Romberg 积分法计算积分 $\int_1^2 e^{\frac{1}{x}} dx$ 的近似值，要求 $|T_m^{(3)} - T_{m-1}^{(3)}|/|T_m^{(3)}| \leqslant 10^{-5}$。

解　在算法(6.22)中，取 $a=1, b=2, f(x)=e^{\frac{1}{x}}$，进行迭代，迭代情况见表 6-5。

表 6-5　例 6 计算结果

$T_m^{(0)}$	$T_m^{(1)}$	$T_m^{(2)}$	$T_m^{(3)}$
2.183 501 550			
2.065 617 795	2.026 323 210		
2.031 892 868	2.020 651 226	2.020 273 094	
2.023 049 868	2.020 102 201	2.020 065 599	2.020 062 306
2.020 808 583	2.020 061 487	2.020 058 773	2.020 058 665

因 $|T_1^{(3)} - T_0^{(3)}|/|T_1^{(3)}| < 10^{-5}$，故得

$$\int_1^2 e^{\frac{1}{x}} dx \approx T_1^{(3)} = 2.020\ 058\ 665$$

6.7 Gauss 型求积公式

6.7.1 一般理论

设要计算下列积分

$$\int_a^b \rho(x) f(x) \mathrm{d}x$$

其中 $\rho(x)$是区间(a,b)上的权函数。

在区间$[a,b]$上取 n 个互异的求积节点 $x_1, x_2, \cdots, x_n$，可形成数值求积公式

$$\int_a^b \rho(x) f(x) \mathrm{d}x \approx \sum_{i=1}^{n} A_i f(x_i) \tag{6.23}$$

其中 $A_i(i=1,2,\cdots,n)$称为求积系数，且与函数 $f(x)$无关。

定义 如果求积公式(6.23)当 $f(x)$为任何次数不高于 m 的多项式时都成为等式，而当 $f(x)$为某 $m+1$ 次多项式时，求积公式(6.23)不能成为等式，则称求积公式(6.23)具有 m 次代数精度。

以求积节点 $x_1, x_2, \cdots, x_n$ 做插值节点，对 $f(x)$进行 Lagrange 插值，并设 $f(x)$在区间$[a,b]$上有 n 阶导数，则有

$$f(x) = \sum_{i=1}^{n} l_i(x) f(x_i) + \frac{f^{(n)}(\xi)}{n!}\omega_n(x)$$

其中当 $x\in[a,b]$时，$\xi=\xi(x)\in(a,b)$，并且

$$\omega_n(x) = (x-x_1)(x-x_2)\cdots(x-x_n)$$

$$l_i(x) = \prod_{\substack{j=1\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j} = \frac{\omega_n(x)}{(x-x_i)\omega'_n(x_i)} \quad (i=1,2,\cdots,n)$$

由此得到插值型求积公式(6.23)，其中求积系数为

$$A_i = \int_a^b \frac{\rho(x)\omega_n(x)}{(x-x_i)\omega'_n(x_i)} \mathrm{d}x \quad (i=1,2,\cdots,n) \tag{6.24}$$

求积公式(6.23)、(6.24)的截断误差为

$$R = \int_a^b \rho(x) f(x) \mathrm{d}x - \sum_{i=1}^{n} A_i f(x_i) = \int_a^b \frac{f^{(n)}(\xi)}{n!}\rho(x)\omega_n(x) \mathrm{d}x \tag{6.25}$$

其中 $\xi=\xi(x)\in(a,b)$。

从截断误差(6.25)看出，无论互异的求积节点 $x_1, x_2, \cdots, x_n$ 在$[a,b]$内如何选取，求积公式(6.23)、(6.24)的代数精度至少是 $n-1$ 次；反之，若求积公式(6.23)的代数精度达到或超过 $n-1$ 次，则它的求积系数必然为式(6.24)。问题是：能否选取适当的节点 $x_1, x_2, \cdots, x_n$，使该求积公式的代数精度提高到 $2n-1$ 次？由于从 $n-1$ 到 $2n-1$ 提高了 n 次，而节点的选择又有 n 个自由度，所以，求积公式(6.23)、(6.24)的代数精度是有可能达到 $2n-1$ 次的。另一方面，不存在这样的节点 $x_i\in[a,b](i=1,2,\cdots,n)$和求积系数 $A_i(i=1,2,\cdots,n)$，使求积公式(6.23)的代数精度达到 $2n$ 次。事实上，只要令

$$\varphi(x) = (x-x_1)^2(x-x_2)^2\cdots(x-x_n)^2$$

其中互异的 $x_1,x_2,\cdots,x_n$ 在$[a,b]$内任取，就有

$$\int_a^b \rho(x)\varphi(x)\mathrm{d}x > 0$$

但对任意的求积系数 $A_i(i=1,2,\cdots,n)$，恒有

$$\sum_{i=1}^{n} A_i\varphi(x_i) = 0$$

可见，求积公式(6.23) 当 $f(x)$为 $2n$ 次多项式 $\varphi(x)$时不可能成为等式。

定义　如果 n 个节点的求积公式(6.23)、(6.24)的代数精度为 $2n-1$ 次，则称它为 Gauss 型求积公式。

n 个节点的所有求积公式中，具有最高代数精度的求积公式是 Gauss 型求积公式。

定理 6.5　设$\{g_k(x)(k=0,1,\cdots)\}$是区间$[a,b]$上带权 $\rho(x)$的正交多项式系，则求积公式(6.23)、(6.24)是 Gauss 型求积公式的充分必要条件是它的求积节点是 n 次正交多项式 $g_n(x)$的n 个零点 $x_i(i=1,2,\cdots,n)$。

证

必要性　由于当 $f(x)$是任何次数不高于 $2n-1$ 的多项式时，求积公式(6.23)成为等式，所以对任意次数不高于 $n-1$ 的多项式 $q(x)$，总有

$$\int_a^b \rho(x)q(x)\omega_n(x)\mathrm{d}x = \sum_{i=1}^{n} A_i q(x_i)\omega_n(x_i) = 0$$

根据定理 5.6 以及正交多项式的唯一性，可知

$$\omega_n(x) \equiv \frac{1}{a_n}g_n(x) \tag{6.26}$$

其中 a_n 是 $g_n(x)$的 x^n 项系数。因此，求积公式(6.23)、(6.24)的求积节点是 $g_n(x)$的零点。

充分性　设求积公式(6.23)、(6.24)的求积节点 $x_1,x_2,\cdots,x_n$ 是多项式 $g_n(x)$的 n 个零点。根据正交多项式的性质，这些节点必互异，且全都在区间(a,b)内。任取次数不高于 $2n-1$ 的多项式 $p(x)$，则 $p(x)$总可表示为

$$p(x) = q(x)\omega_n(x) + r(x)$$

其中 $q(x)$和 $r(x)$都是次数不高于 $n-1$ 的多项式，并且 $p(x_i)=r(x_i)(i=1,2,\cdots,n)$。于是有

$$\int_a^b \rho(x)p(x)\mathrm{d}x = \int_a^b \rho(x)q(x)\omega_n(x)\mathrm{d}x + \int_a^b \rho(x)r(x)\mathrm{d}x =$$

$$\frac{1}{a_n}\int_a^b \rho(x)q(x)g_n(x)\mathrm{d}x + \sum_{i=1}^{n} A_i r(x_i) = \sum_{i=1}^{n} A_i p(x_i)$$

由此可知，求积公式(6.23)、(6.24)是 Gauss 型求积公式。

证毕。

由式(6.26)可知，Gauss 型求积公式(6.23)、(6.24)的求积系数可表示为

$$A_i = \int_a^b \frac{\rho(x)g_n(x)}{(x-x_i)g'_n(x_i)}\mathrm{d}x \quad (i=1,2,\cdots,n) \tag{6.27}$$

定理 6.6　设 $f(x)$在区间$[a,b]$上有 $2n$ 阶连续导数，则 Gauss 型求积公式(6.23)、(6.27)的截断误差为

$$R = \frac{f^{(2n)}(\eta)}{a_n^2(2n)!}\int_a^b \rho(x)g_n^2(x)\mathrm{d}x \tag{6.28}$$

其中 $\eta\in(a,b)$，a_n 是正交多项式 $g_n(x)$ 的最高次项系数。

证 以求积节点 $x_1,x_2,\cdots,x_n$ 为插值节点，可作出满足插值条件

$$H_{2n-1}(x_i)=f(x_i),\quad H'_{2n-1}(x_i)=f'(x_i)\quad(i=1,2,\cdots,n)$$

的 $2n-1$ 次 Hermite 插值多项式 $H_{2n-1}(x)$。根据定理 5.3，有

$$f(x)=H_{2n-1}(x)+\frac{f^{(2n)}(\xi)}{(2n)!}\omega_n^2(x)\tag{6.29}$$

其中当 $x\in[a,b]$ 时，$\xi=\xi(x)\in(a,b)$。

由式(6.29)可得

$$\int_a^b\rho(x)f(x)\mathrm{d}x=\int_a^b\rho(x)H_{2n-1}(x)\mathrm{d}x+\int_a^b\rho(x)\frac{f^{(2n)}(\xi)}{(2n)!}\omega_n^2(x)\mathrm{d}x$$

因求积公式(6.23)、(6.27)是 Gauss 型的，故有

$$\int_a^b\rho(x)H_{2n-1}(x)\mathrm{d}x=\sum_{i=1}^nA_iH_{2n-1}(x_i)=\sum_{i=1}^nA_if(x_i)$$

又因 $f^{(2n)}(x)\in C[a,b]$ 以及 $\rho(x)\omega_n^2(x)$ 在 (a,b) 上不变号，根据积分中值定理，可知

$$\int_a^b\rho(x)\frac{f^{(2n)}(\xi)}{(2n)!}\omega_n^2(x)\mathrm{d}x=\frac{f^{(2n)}(\eta)}{(2n)!}\int_a^b\rho(x)\omega_n^2(x)\mathrm{d}x=$$

$$\frac{f^{(2n)}(\eta)}{a_n^2(2n)!}\int_a^b\rho(x)g_n^2(x)\mathrm{d}x,\quad\eta\in(a,b)$$

于是有

$$R=\int_a^b\rho(x)f(x)\mathrm{d}x-\sum_{i=1}^nA_if(x_i)=\frac{f^{(2n)}(\eta)}{a_n^2(2n)!}\int_a^b\rho(x)g_n^2(x)\mathrm{d}x,\quad\eta\in(a,b)$$

证毕。

定理 6.7 设求积公式(6.23)、(6.24)是 Gauss 型求积公式，则它的求积系数 A_i 满足

(1) $A_i>0(i=1,2,\cdots,n)$；

(2) $\displaystyle\sum_{i=1}^nA_i=\int_a^b\rho(x)\mathrm{d}x$。

证

(1) 令

$$\varphi_k(x)=\prod_{\substack{j=1\\j\neq k}}^n(x-x_j)^2\quad(k=1,2,\cdots,n)$$

其中 $x_i(i=1,2,\cdots,n)$ 是求积节点。由于 $\varphi_k(x)$ 是 $2n-2$ 次多项式，故有

$$\int_a^b\rho(x)\varphi_k(x)\mathrm{d}x=\sum_{i=1}^nA_i\varphi_k(x_i)=A_k\varphi_k(x_k)\quad(k=1,2,\cdots,n)$$

因 $\rho(x)\varphi_k(x)$ 在 (a,b) 内非负且不恒为零，故

$$\int_a^b\rho(x)\varphi_k(x)\mathrm{d}x>0\quad(k=1,2,\cdots,n)$$

又因 $\varphi_k(x_k)>0$，从而必有 $A_k>0\quad(k=1,2,\cdots,n)$。

(2) 因 $n\geqslant1$，故任何 Gauss 型求积公式当 $f(x)$ 恒为常数时都会成为等式。取 $f(x)\equiv1$，就得

$$\int_a^b\rho(x)\mathrm{d}x=\sum_{i=1}^nA_i$$

证毕。

根据定理 6.7 可知，Gauss 型求积公式的求积系数满足

$$\sum_{i=1}^{n} |A_i| = \int_a^b \rho(x)\mathrm{d}x \quad （常数）$$

故 Gauss 型求积公式具有数值稳定性。

定理 6.8　设求积公式(6.23)、(6.24)是 Gauss 型求积公式。若函数 $f(x)\in C[a,b]$，则有

$$\lim_{n\to\infty}\sum_{i=1}^{n} A_i f(x_i) = \int_a^b \rho(x)f(x)\mathrm{d}x$$

证　因 $f(x)\in C[a,b]$，据 Weierstrass（魏尔斯托拉斯）定理（参阅文献[2]第 159 页），对任意给定的正数 ε，必存在多项式 $p(x)$（设为 m 次）满足不等式

$$|f(x)-p(x)| < \frac{\varepsilon}{2}\left[\int_a^b \rho(x)\mathrm{d}x\right]^{-1}, \quad x\in[a,b]$$

记

$$Q_n(f) = \sum_{i=1}^{n} A_i f(x_i)$$

则有

$$\int_a^b \rho(x)f(x)\mathrm{d}x - Q_n(f) = \int_a^b \rho(x)[f(x)-p(x)]\mathrm{d}x + \int_a^b \rho(x)p(x)\mathrm{d}x - Q_n(p) + Q_n(p) - Q_n(f)$$

当 $n > \frac{m+1}{2}$ 时，$\int_a^b \rho(x)p(x)\mathrm{d}x - Q_n(p) = 0$；又因 $A_i > 0$，故有

$$|Q_n(p) - Q_n(f)| \leqslant \sum_{i=1}^{n} A_i |p(x_i) - f(x_i)|$$

于是，存在 $N \geqslant \frac{m+1}{2}$，当 $n > N$ 时恒有

$$\begin{aligned}\left|\int_a^b \rho(x)f(x)\mathrm{d}x - Q_n(f)\right| &\leqslant \int_a^b \rho(x)|f(x)-p(x)|\mathrm{d}x + \\ &\sum_{i=1}^{n} A_i |p(x_i)-f(x_i)| \leqslant \\ &\frac{\varepsilon}{2}\left[\int_a^b \rho(x)\mathrm{d}x\right]^{-1}\int_a^b \rho(x)\mathrm{d}x + \\ &\frac{\varepsilon}{2}\left[\int_a^b \rho(x)\mathrm{d}x\right]^{-1}\sum_{i=1}^{n} A_i = \\ &\frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\end{aligned}$$

即 $\lim_{n\to\infty} Q_n(f) = \int_a^b \rho(x)f(x)\mathrm{d}x$ 成立。

证毕。

由于正交多项式随区间和权函数的不同而不同，因此有不同类型的 Gauss 型求积公式。

理论上，凡是给定了积分区间$[a,b]$和在(a,b)上的权函数 $\rho(x)$，并给定节点个数 n，就可

以构造形如式(6.23)的 Gauss 型求积公式，方法如下。

首先求出在$[a,b]$上带权$\rho(x)$的正交多项式系$\{g_k(x)\}$中的n次多项式$g_n(x)$；然后求出方程$g_n(x)=0$的n个根$x_1,x_2,\cdots,x_n$，这就是求积节点；最后按公式(6.27)即下列公式

$$A_i=\int_a^b \rho(x)\left(\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\right)\mathrm{d}x \quad (i=1,2,\cdots,n)$$

求出求积系数$A_i(i=1,2,\cdots,n)$。也可根据 Gauss 型求积公式的代数精度形成一个关于$A_i(i=1,2,\cdots,n)$的n元线性方程组，求解此方程组就得到求积系数$A_i(i=1,2,\cdots,n)$。

例 7 试构造形如

$$\int_{-1}^{1}x^2 f(x)\mathrm{d}x\approx\sum_{i=1}^{3}A_i f(x_i)$$

的 Gauss 型求积公式。

解 在 5.5.1 小节已求出在区间$[-1,1]$上带权$\rho(x)=x^2$的正交多项式组

$$g_0(x)\equiv 1,\quad g_1(x)=x,\quad g_2(x)=x^2-\frac{3}{5},\quad g_3(x)=x^3-\frac{5}{7}x$$

由方程$g_3(x)=0$解出求积节点

$$x_1=-\sqrt{\frac{5}{7}},\quad x_2=0,\quad x_3=\sqrt{\frac{5}{7}}$$

计算求积系数

$$A_1=\int_{-1}^{1}x^2\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}\mathrm{d}x=\frac{7}{25}$$

$$A_2=\int_{-1}^{1}x^2\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}\mathrm{d}x=\frac{8}{75}$$

由$A_1+A_2+A_3=\int_{-1}^{1}x^2\mathrm{d}x=\frac{2}{3}$得到

$$A_3=\frac{2}{3}-A_1-A_2=\frac{7}{25}$$

所要的 Gauss 型求积公式为

$$\int_{-1}^{1}x^2 f(x)\mathrm{d}x\approx\frac{1}{25}\left[7f\left(-\sqrt{\frac{5}{7}}\right)+\frac{8}{3}f(0)+7f\left(\sqrt{\frac{5}{7}}\right)\right]$$

下面介绍几种常用的 Gauss 型求积公式。

6.7.2 几种 Gauss 型求积公式

1. Gauss - Legendre 求积公式

给定权函数$\rho(x)\equiv 1$和积分区间$[-1,1]$，由 5.5 节知，其相应的正交多项式是 Legendre 多项式

$$L_n(x)=\frac{1}{2^n n!}\cdot\frac{\mathrm{d}^n}{\mathrm{d}x^n}[(x^2-1)^n]$$

取$L_n(x)$的零点作求积节点所形成的求积公式

$$\int_{-1}^{1}f(x)\mathrm{d}x\approx\sum_{i=1}^{n}A_i f(x_i) \tag{6.30}$$

称为 Gauss - Legendre 求积公式，其中

$$A_i = \int_{-1}^{1} \frac{L_n(x)}{(x-x_i)L'_n(x_i)}\mathrm{d}x \quad (i=1,2,\cdots,n)$$

但直接由上式计算 A_i 很困难，下面用间接的方法求出 A_i，为此，考察积分

$$s_i = \int_{-1}^{1} \frac{L_n(x)L'_n(x)}{x-x_i}\mathrm{d}x$$

因被积函数是 $2n-2$ 次多项式，故由 Gauss - Legendre 求积公式(6.30)，有

$$s_i = \sum_{j=1}^{n} A_j \frac{L_n(x_j)L'_n(x_j)}{x_j - x_i} = A_i[L'_n(x_i)]^2$$

另一方面，按照分部积分法，令 $u=\dfrac{L_n(x)}{x-x_i}$，则

$$s_i = \left.\frac{L_n^2(x)}{x-x_i}\right|_{-1}^{1} - \int_{-1}^{1} L_n(x)u'\mathrm{d}x$$

因 u' 是次数小于 $n-1$ 的多项式，故上式右端的积分等于零。又根据 Legendre 多项式的性质，可知 $L_n^2(\pm 1)=1$，因此有

$$s_i = \frac{1}{1-x_i} + \frac{1}{1+x_i} = \frac{2}{1-x_i^2}$$

最后求得 Gauss - Legendre 求积公式(6.30)的求积系数为

$$A_i = \frac{2}{(1-x_i^2)[L'_n(x_i)]^2} \quad (i=1,2,\cdots,n)$$

根据式(6.28)和 n 次 Legendre 多项式的最高次项系数 $a_n = \dfrac{(2n)!}{2^n(n!)^2}$ 可知，Gauss - Legendre求积公式(6.30)的截断误差为

$$R = \frac{f^{(2n)}(\eta)}{a_n^2(2n)!}\int_{-1}^{1} L_n^2(x)\mathrm{d}x = \frac{f^{(2n)}(\eta)}{(2n)!} \cdot \frac{2^{2n}(n!)^4}{[(2n)!]^2} \cdot \frac{2}{2n+1}$$

其中 $\eta \in (-1,1)$。

Gauss - Legendre 求积公式的节点 x_i 及求积系数 A_i 见表6 - 6。

例 8　分别用三点 Simpson 公式和三点 Gauss - Legendre 求积公式计算积分

$$\int_{-1}^{1} \sqrt{x+1}\mathrm{d}x$$

解　用三点 Simpson 公式计算，得

$$\int_{-1}^{1} \sqrt{x+1}\mathrm{d}x \approx \frac{2}{6}(\sqrt{-1+1} + 4\sqrt{0+1} + \sqrt{1+1}) = 1.804\ 737\ 854$$

用三点 Gauss - Legendre 求积公式计算，得

$$\begin{aligned}\int_{-1}^{1} \sqrt{x+1}\mathrm{d}x \approx\ & 0.555\ 555\ 555\ 6\sqrt{-0.774\ 596\ 669\ 2+1} + \\ & 0.888\ 888\ 888\ 9\sqrt{0+1} + \\ & 0.555\ 555\ 555\ 6\sqrt{0.774\ 596\ 669\ 2+1} = \\ & 1.892\ 725\ 829\end{aligned}$$

精确值为 1.885 618 083。可见，在节点数相同的情况下，用 Gauss 型求积公式计算，结果的精确度高于用 Newton - Cotes 求积公式计算。

表 6-6 Gauss-Legendre 求积节点和求积系数

n	x_i	A_i	n	x_i	A_i
1	0	2	7	±0.949 107 912 3	0.129 484 966 2
2	±0.577 350 269 2	1		±0.741 531 185 6	0.279 705 391 5
3	±0.774 596 669 2	0.555 555 555 6		±0.405 845 151 4	0.381 830 050 5
	0	0.888 888 888 9		0	0.417 959 183 7
4	±0.861 136 311 6	0.347 854 845 1	8	±0.960 289 856 5	0.101 228 536 3
	±0.339 981 043 6	0.652 145 154 9		±0.796 666 477 4	0.222 381 034 5
5	±0.906 179 845 9	0.236 926 885 1		±0.525 532 409 9	0.313 706 645 9
	±0.538 469 310 1	0.478 628 670 5		±0.183 434 642 5	0.362 683 783 4
	0	0.568 888 888 9	9	±0.968 160 239 5	0.081 274 388 4
6	±0.932 469 514 2	0.171 324 492 4		±0.836 031 107 3	0.180 648 160 7
	±0.661 209 386 5	0.360 761 573 0		±0.613 371 432 7	0.260 610 696 4
	±0.238 619 186 1	0.467 913 934 6		±0.324 253 423 4	0.312 347 077 0
				0	0.330 239 355 0

如果积分区间是$[a,b]$,那么通过变量置换

$$x=\frac{a+b}{2}+\frac{b-a}{2}t$$

就可以使用 Gauss-Legendre 求积公式计算积分

$$\int_a^b f(x)\mathrm{d}x=\frac{b-a}{2}\int_{-1}^{1}f\left(\frac{a+b}{2}+\frac{b-a}{2}t\right)\mathrm{d}t$$

例 9 用四点 Gauss-Legendre 求积公式计算积分

$$\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x$$

解 令 $x=\frac{1}{2}(t+3)$,则

$$\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x=\frac{1}{2}\int_{-1}^{1}\mathrm{e}^{\frac{2}{t+3}}\mathrm{d}t$$

记 $\varphi(t)=\mathrm{e}^{\frac{2}{t+3}}$,则由

$$\varphi(t_1)=\varphi(-0.861\ 136\ 311\ 6)=2.547\ 406\ 932$$
$$\varphi(t_2)=\varphi(-0.339\ 981\ 043\ 6)=2.120\ 971\ 718$$
$$\varphi(t_3)=\varphi(0.339\ 981\ 043\ 6)=1.819\ 944\ 113$$
$$\varphi(t_4)=\varphi(0.861\ 136\ 311\ 6)=1.678\ 637\ 128$$

得

$$\int_1^2 \mathrm{e}^{\frac{1}{x}}\mathrm{d}x\approx\frac{1}{2}\{0.347\ 854\ 845\ 1[\varphi(t_1)+\varphi(t_4)]+0.652\ 145\ 154\ 9[\varphi(t_2)+\varphi(t_3)]\}=2.020\ 049\ 534$$

2. Gauss-Laguerre 求积公式

当权函数 $\rho(x)=\mathrm{e}^{-x}$,积分区间为$[0,\infty)$时,取区间$[0,\infty)$上带权 $\rho(x)=\mathrm{e}^{-x}$的 n 次正交

多项式——Laguerre 多项式

$$U_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x})$$

的零点作求积节点所形成的求积公式

$$\int_0^{\infty} e^{-x} f(x) dx \approx \sum_{i=1}^{n} A_i f(x_i) \tag{6.31}$$

称为 Gauss - Laguerre 求积公式,其中

$$A_i = \int_0^{\infty} e^{-x} \frac{U_n(x)}{(x - x_i)U'_n(x_i)} dx \quad (i = 1,2,\cdots,n)$$

用类似于 Gauss - Legendre 求积公式的间接方法,可求出

$$A_i = \frac{(n!)^2}{x_i[U'_n(x_i)]^2} \quad (i = 1,2,\cdots,n)$$

根据式(6. 28)和 n 次 Laguerre 多项式的最高次项系数 $a_n = (-1)^n$,可知 Gauss - Laguerre 求积公式(6. 31)的截断误差为

$$R = \frac{f^{(2n)}(\eta)}{a_n^2(2n)!}\int_0^{\infty} e^{-x} U_n^2(x) dx = \frac{(n!)^2}{(2n)!} f^{(2n)}(\eta)$$

其中 $\eta \in (0,\infty)$。

Gauss - Laguerre 求积公式的节点 x_i 及求积系数 A_i 见表6 - 7。

表 6 - 7 Gauss - Laguerre 求积节点和求积系数

n	x_i	A_i	n	x_i	A_i
1	1	1	5	0. 263 560 319 7	0. 521 755 610 6
2	0. 585 786 437 6	0. 853 553 390 6		1. 413 403 059 1	0. 398 666 811 1
	3. 414 213 562 4	0. 146 446 609 4		3. 596 425 771 0	0. 075 942 449 7
3	0. 415 774 556 8	0. 711 093 009 9		7. 085 810 005 9	0. 003 611 758 7
	2. 294 280 360 3	0. 278 517 733 6		12. 640 800 844	0. 000 023 370 0
	6. 289 945 082 9	0. 010 389 256 5	6	0. 222 846 604 2	0. 458 964 674 0
4	0. 322 547 689 6	0. 603 154 104 3		1. 188 932 101 7	0. 417 000 830 8
	1. 745 761 101 2	0. 357 418 692 4		2. 992 736 326 1	0. 113 373 382 1
	4. 536 620 296 9	0. 038 887 908 5		5. 775 143 569 1	0. 010 399 197 5
	9. 395 070 912 3	0. 000 539 294 7		9. 837 467 418 4	0. 000 261 017 2
				15. 982 873 981	0. 000 000 898 5

3. Gauss - Hermite 求积公式

当权函数 $\rho(x) = e^{-x^2}$,积分区间为$(-\infty,\infty)$时,得求积公式

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{i=1}^{n} A_i f(x_i) \tag{6.32}$$

称式(6. 32)为 Gauss - Hermite 求积公式,其中 x_i 是区间$(-\infty,\infty)$上带权 $\rho(x) = e^{-x^2}$ 的 n 次正交多项式——Hermite 多项式

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

的零点，求积系数为

$$A_i = \int_{-\infty}^{\infty} e^{-x^2} \frac{H_n(x)}{(x-x_i)H'_n(x_i)} dx = \frac{2^{n+1} n! \sqrt{\pi}}{[H'_n(x_i)]^2} \quad (i=1,2,\cdots,n)$$

根据式(6.28)和 Hermite 多项式 $H_n(x)$ 的最高次项系数 $a_n = 2^n$，可知 Gauss - Hermite 求积公式(6.32)的截断误差为

$$R = \frac{f^{(2n)}(\eta)}{a_n^2 (2n)!} \int_{-\infty}^{\infty} e^{-x^2} H_n^2(x) dx = \frac{n! \sqrt{\pi}}{2^n (2n)!} f^{(2n)}(\eta)$$

其中 $\eta \in (-\infty, \infty)$。

Gauss - Hermite 求积公式的节点 x_i 及求积系数 A_i 见表6 - 8。

表 6 - 8　Gauss - Hermite 求积节点和求积系数

n	x_i	A_i	n	x_i	A_i
1	0	1.772 453 850 0		±1.335 849 074 0	0.157 067 320 3
2	±0.707 106 781 2	0.886 226 925 5		±0.436 077 411 9	0.724 629 595 2
3	±1.224 744 871 4	0.295 408 975 2	7	±2.651 961 356 8	0.000 971 781 245
	0	1.181 635 900 6		±1.673 551 628 8	0.054 515 582 82
4	±1.650 680 123 9	0.081 312 835 5		±0.816 287 882 9	0.425 607 252 6
	±0.524 647 623 3	0.804 914 090 0		0	0.810 264 617 6
5	±2.020 182 870 5	0.019 953 242 1	8	±2.930 637 420 3	0.000 199 604 07
	±0.958 572 464 6	0.393 619 323 2		±1.981 656 756 7	0.017 077 983 01
	0	0.945 308 720 5		±1.157 193 712 4	0.207 802 325 8
6	±2.350 604 973 7	0.004 530 009 9		±0.381 186 990 2	0.661 147 012 6

4. Gauss - Chebyshev 求积公式

当权函数 $\rho(x) = \frac{1}{\sqrt{1-x^2}}$，积分区间为$[-1,1]$，得求积公式

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} dx \approx \sum_{i=1}^{n} A_i f(x_i) \tag{6.33}$$

称式(6.33)为 Gauss - Chebyshev 求积公式，其中

$$x_i = \cos \frac{2(n-i)+1}{2n} \pi \quad (i=1,2,\cdots,n)$$

是 n 次 Chebyshev 多项式

$$T_n(x) = \cos(n \arccos x)$$

的零点，求积系数为

$$A_i = \int_{-1}^{1} \frac{T_n(x) dx}{\sqrt{1-x^2}(x-x_i)T'_n(x_i)} = \frac{\pi}{n} \quad (i=1,2,\cdots,n)$$

根据式(6.28)和 Chebyshev 多项式 $T_n(x)$ 的最高次项系数 $a_n = 2^{n-1}$，可知 Gauss -

Chebyshev求积公式(6.33)的截断误差为

$$R=\frac{f^{(2n)}(\eta)}{a_n^2(2n)!}\int_{-1}^{1}\frac{T_n^2(x)}{\sqrt{1-x^2}}\mathrm{d}x=\frac{2\pi}{2^{2n}(2n)!}f^{(2n)}(\eta)$$

其中 $\eta\in(-1,1)$。

例 10　计算积分 $\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x$，其中 $f(x)$是 m 次多项式。

解　使用 Gauss－Chebyshev 求积公式(6.33)计算。当 m 为奇数时，由于 $f(x)$的 $m+1$ 阶导数恒为零，所以在式(6.33)中取$n=\frac{m+1}{2}$可使截断误差为零，于是得

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x=\frac{2\pi}{m+1}\sum_{i=1}^{\frac{m+1}{2}}f\left(\cos\frac{m+2-2i}{m+1}\pi\right)$$

当 m 为偶数时，应取 $n=\frac{m+2}{2}$才能使截断误差为零，此时得

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x=\frac{2\pi}{m+2}\sum_{i=1}^{\frac{m+2}{2}}f\left(\cos\frac{m+3-2i}{m+2}\pi\right)$$

例 11　计算积分 $\int_{-1}^{1}\sqrt{\frac{2+x}{1-x^2}}\mathrm{d}x$。

解　选用 $n=3$ 的 Gauss－Chebyshev 求积公式(6.33)计算。这时

$$x_1=\cos\frac{5}{6}\pi=-0.866\ 025\ 403$$

$$x_2=\cos\frac{3}{6}\pi=0$$

$$x_3=\cos\frac{1}{6}\pi=0.866\ 025\ 403$$

$$A_i=\frac{\pi}{3}\quad(i=1,2,3)$$

于是有

$$\int_{-1}^{1}\sqrt{\frac{2+x}{1-x^2}}\mathrm{d}x\approx\frac{\pi}{3}(\sqrt{2+x_1}+\sqrt{2+x_2}+\sqrt{2+x_3})=4.368\ 939\ 556$$

Gauss 型求积公式使用少节点可得高精度的结果，是其明显优点。例如，计算积分

$$I=\int_0^1\frac{\mathrm{d}x}{1+x}$$

它的精确值(取八位有效数字)为

$$I=0.693\ 147\ 18$$

使用节点数为 129 的复化 Simpson 公式计算，得

$$I\approx0.693\ 146\ 70$$

使用节点数为 10 的 Gauss－Legendre 求积公式计算，得

$$I\approx0.693\ 147\ 10$$

可见，后者只用 10 个节点所得结果比前者用 129 个节点所得结果还要准确。

Gauss 型求积公式的另一明显优点是能计算许多广义积分，而这些广义积分如果使用

Newton - Cotes 求积公式是很难处理的。

Gauss 型求积公式也有明显的缺点，就是节点和求积系数需要查表，并且当节点数目 n 增加时，原来的节点几乎都不能用，先前计算的被积函数值不能重复使用，这将造成不必要的浪费。但是，在很多情况下，由于 Gauss 型求积公式的快速收敛性，上述的低效率是无关紧要的。

6.8 二重积分的数值求积法

这里只介绍一种数值求积法，这种方法是把二重积分化为二次积分，然后利用定积分计算中的梯形公式或者 Simpson 公式计算二次积分，从而计算出二重积分的近似值。

6.8.1 矩形域上的二重积分

设有二重积分

$$I(f) = \iint_D f(x,y)\mathrm{d}x\mathrm{d}y$$

其中积分域为矩形域 $D=\{(x,y)\,|\,a\leqslant x\leqslant b, c\leqslant y\leqslant d\}$。$I(f)$可以化为二次积分

$$I(f) = \int_c^d \mathrm{d}y\int_a^b f(x,y)\mathrm{d}x$$

1. 复化梯形公式

利用梯形公式(6.8)，得

$$\int_a^b f(x,y)\mathrm{d}x \approx \frac{b-a}{2}[f(a,y)+f(b,y)]$$

$$\begin{aligned} I(f) \approx \frac{b-a}{2}\left[\int_c^d f(a,y)\mathrm{d}y + \int_c^d f(b,y)\mathrm{d}y\right] = \\ \frac{(b-a)(d-c)}{4}[f(a,c)+f(a,d)+ \\ f(b,c)+f(b,d)] \end{aligned} \tag{6.34}$$

式(6.34)称为计算二重积分 $I(f)$的梯形公式。

取

$$x_i = a+ih \quad (i=0,1,\cdots,m) \quad h=\frac{b-a}{m}$$

$$y_j = c+j\tau \quad (j=0,1,\cdots,n) \quad \tau=\frac{d-c}{n}$$

则直线族 $x=x_i(i=0,1,\cdots,m)$和直线族 $y=y_j(j=0,1,\cdots,n)$把域 D 分割成 $m\times n$ 个子矩形域

$$D_{ij} = \{(x,y)\mid x_i\leqslant x\leqslant x_{i+1}, y_j\leqslant y\leqslant y_{j+1}\}$$
$$(i=0,1,\cdots,m-1; j=0,1,\cdots,n-1)$$

两直线族的交点(x_i,y_j) $(i=0,1,\cdots,m; j=0,1,\cdots,n)$称为求积节点。记 $f_{ij}=f(x_i,y_j)$，则由梯形公式(6.34)可得

$$I(f)=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\iint\limits_{D_{ij}}f(x,y)\mathrm{d}x\mathrm{d}y\approx$$
$$\frac{h\tau}{4}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(f_{ij}+f_{i,j+1}+f_{i+1,j}+f_{i+1,j+1})=\frac{h\tau}{4}\sum_{i=0}^{m}\sum_{j=0}^{n}\lambda_{ij}f_{ij} \tag{6.35}$$

其中

$$\begin{cases}\lambda_{00}=\lambda_{0n}=\lambda_{m0}=\lambda_{mn}=1\\ \lambda_{i0}=\lambda_{in}=2\quad(i=1,2,\cdots,m-1)\\ \lambda_{0j}=\lambda_{mj}=2\quad(j=1,2,\cdots,n-1)\\ \lambda_{ij}=4\quad(i=1,2,\cdots,m-1;j=1,2,\cdots,n-1)\end{cases}$$

式(6.35)称为计算二重积分 $I(f)$的复化梯形公式。

2. 复化 Simpson 公式

利用 Simpson 公式(6.10)，得

$$\int_a^b f(x,y)\mathrm{d}x\approx\frac{b-a}{6}\left[f(a,y)+4f\left(\frac{a+b}{2},y\right)+f(b,y)\right]$$
$$I(f)\approx\frac{b-a}{6}\int_c^d\left[f(a,y)+4f\left(\frac{a+b}{2},y\right)+f(b,y)\right]\mathrm{d}y\approx$$
$$\frac{(b-a)(d-c)}{36}\left\{f(a,c)+f(b,c)+f(a,d)+f(b,d)+\right.$$
$$4\left[f\left(\frac{a+b}{2},c\right)+f\left(\frac{a+b}{2},d\right)+f\left(a,\frac{c+d}{2}\right)+f\left(b,\frac{c+d}{2}\right)\right]+$$
$$\left.16f\left(\frac{a+b}{2},\frac{c+d}{2}\right)\right\} \tag{6.36}$$

式(6.36)称为计算二重积分 $I(f)$的 Simpson 公式。

取

$$x_k=a+kh\quad(k=0,1,\cdots,2m)\quad h=\frac{b-a}{2m}$$
$$y_l=c+l\tau\quad(l=0,1,\cdots,2n)\quad \tau=\frac{d-c}{2n}$$

得求积节点$(x_k,y_l)(k=0,1,\cdots,2m;l=0,1,\cdots,2n)$以及子矩形域

$$D_{ij}=\{(x,y)\mid x_{2i}\leqslant x\leqslant x_{2i+2},y_{2j}\leqslant y\leqslant y_{2j+2}\}$$
$$(i=0,1,\cdots,m-1;j=0,1,\cdots,n-1)$$

利用 Simpson 公式(6.36)得

$$I(f)=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\iint\limits_{D_{ij}}f(x,y)\mathrm{d}x\mathrm{d}y\approx$$
$$\frac{(2h)(2\tau)}{36}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[f_{2i,2j}+f_{2i+2,2j}+f_{2i,2j+2}+$$
$$f_{2i+2,2j+2}+4(f_{2i+1,2j}+f_{2i+1,2j+2}+f_{2i,2j+1}+$$
$$f_{2i+2,2j+1})+16f_{2i+1,2j+1}]=\frac{h\tau}{9}\sum_{i=0}^{2m}\sum_{j=0}^{2n}\lambda_{ij}f_{ij} \tag{6.37}$$

其中

$$\begin{cases}\lambda_{00}=\lambda_{0,2n}=\lambda_{2m,0}=\lambda_{2m,2n}=1\\ \lambda_{i0}=\lambda_{i,2n}=4\quad(i=1,3,\cdots,2m-1)\\ \lambda_{0j}=\lambda_{2m,j}=4\quad(j=1,3,\cdots,2n-1)\\ \lambda_{ij}=4\quad(i=2,4,\cdots,2m-2;j=2,4,\cdots,2n-2)\\ \lambda_{i0}=\lambda_{i,2n}=2\quad(i=2,4,\cdots,2m-2)\\ \lambda_{0j}=\lambda_{2m,j}=2\quad(j=2,4,\cdots,2n-2)\\ \lambda_{ij}=8\quad(i=1,3,\cdots,2m-1;j=2,4,\cdots,2n-2)\\ \lambda_{ij}=8\quad(i=2,4,\cdots,2m-2;j=1,3,\cdots,2n-1)\\ \lambda_{ij}=16\quad(i=1,3,\cdots,2m-1;j=1,3,\cdots,2n-1)\end{cases}$$

式(6.37)称为计算二重积分 $I(f)$的复化 Simpson 公式。

6.8.2 一般区域上的二重积分

设 $I(f)=\iint\limits_D f(x,y)\mathrm{d}x\mathrm{d}y$ 的积分区域 D 为

$$D=\{(x,y)\mid u(x)\leqslant y\leqslant v(x),a\leqslant x\leqslant b\}$$

其中 $u(x),v(x)$都是区间$[a,b]$上的连续函数。作矩形 $R=\{(x,y)\mid a\leqslant x\leqslant b,c\leqslant y\leqslant d\}$，使得 $D\subset R$(见图 6-1)。

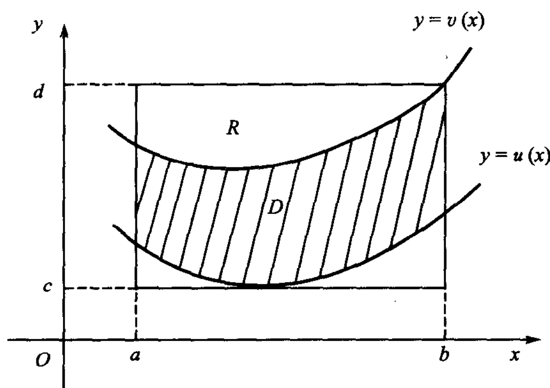


图 6-1 矩形域 R 包含着积分区域 D

令

$$F(x,y)=\begin{cases}f(x,y), & (x,y)\in D\\ 0, & (x,y)\in R\backslash D\end{cases}$$

则由公式(6.37)得

$$I(f)=\iint\limits_R F(x,y)\mathrm{d}x\mathrm{d}y\approx\frac{h\tau}{9}\sum_{i=0}^{2m}\sum_{j=0}^{2n}\lambda_{ij}F_{ij}$$

其中

$$F_{ij}=\begin{cases}f(x_i,y_j), & u(x_i)\leqslant y_j\leqslant v(x_i)\\ 0, & \text{其他}\end{cases}$$

习　题

1. 记 $I(f)=\int_{-1}^{1}f(x)\mathrm{d}x$，求下列求积公式的代数精度：

(1)　$I(f)\approx\frac{2}{3}[f(-1)+f(0)+f(1)]$；

(2)　$I(f)\approx\frac{1}{3}[f(-1)+4f(0)+f(1)]$。

2. 给定求积公式

$$\int_{-2h}^{2h}f(x)\mathrm{d}x\approx\lambda_0f(-h)+\lambda_1f(0)+\lambda_2f(h)$$

其中 $h>0$，试确定 $\lambda_0,\lambda_1,\lambda_2$，使此求积公式的代数精度尽量高。

3. 试确定求积节点 x_0,x_1，使求积公式

$$\int_{1}^{1}f(x)\mathrm{d}x\approx f(x_0)+f(x_1)$$

具有尽可能高的代数精度。

4. 试以 $x_0=-h,x_1=0,x_2=h(0<h<1)$ 为求积节点，推出计算积分 $\int_{-1}^{1}f(x)\mathrm{d}x$ 的插值型求积公式及其截断误差；并确定 h 的值，使此求积公式具有尽可能高的代数精度。

5. 试证：$n+1$ 个节点的求积公式如果具有 n 次或大于 n 次的代数精度，则它是插值型求积公式。

6. 试推出下列两个求积公式的截断误差表达式，并判断其代数精度：

(1)　$\int_a^b f(x)\mathrm{d}x\approx(b-a)f(a)$；

(2)　$\int_a^b f(x)\mathrm{d}x\approx(b-a)f\left(\frac{a+b}{2}\right)$。

7. 试以 $x_0=0,x_1=h,x_2=2h$ 为求积节点，推出计算积分 $\int_0^{3h}f(x)\mathrm{d}x$ 的插值型求积公式，并利用 $f(x)$ 的 Taylor 级数展开式证明此求积公式的截断误差为

$$R=\frac{3}{8}h^4f'''(0)+O(h^5)$$

8. 试分别用 $n=6$ 的复化梯形公式和 $m=3$ 的复化 Simpson 公式计算积分 $\int_1^2\frac{\mathrm{e}^{-x}}{x}\mathrm{d}x$ 的近似值，并问：所得的近似值至少有几位有效数字？

9. 若用 $n+1$ 个节点的复化梯形公式计算积分 $\int_0^1\mathrm{e}^{-x^2}\mathrm{d}x$ 的近似值，则 n 取何值时能保证计算结果有四位有效数字（假定计算过程无舍入误差）？

10. 设 T_m 是 2^m+1 个节点的复化梯形值，试证：

$$S_m=\frac{4T_{m+1}-T_m}{3}$$

就是 $2^{m+1}+1$ 个节点的复化 Simpson 值。

11. 试用区间逐次分半的复化梯形公式计算积分 $\int_0^1 e^{-x^2}\,dx$ 的近似值 T_m，要求 $|T_m - T_{m-1}|/|T_m| \leqslant 10^{-4}$。

12. 试用 Romberg 积分法计算积分 $\int_0^1 e^{-x^2}\,dx$ 的近似值 $T_m^{(3)}$，要求 $|T_m^{(3)} - T_{m-1}^{(3)}|/|T_m^{(3)}| \leqslant 10^{-4}$。

13. 试叙述 $n \geqslant 2$ 的复化梯形公式和 $m \geqslant 2$ 的复化 Simpson 公式都不是插值型求积公式的理由。

14. 下列求积公式中，哪一个属于 Gauss 型求积公式：

(1) $\int_{-1}^1 f(x)\,dx \approx \frac{1}{3}[f(-1) + 4f(0) + f(1)]$；

(2) $\int_{-1}^1 f(x)\,dx \approx \frac{1}{9}[5f(-\sqrt{0.6}) + 8f(0) + 5f(\sqrt{0.6})]$；

(3) $\int_0^2 f(x)\,dx \approx f\left(1 - \frac{1}{\sqrt{3}}\right) + f\left(1 + \frac{1}{\sqrt{3}}\right)$。

15. 利用 $n=4$ 的 Gauss - Legendre 求积公式计算下列积分：

(1) $\int_{-1}^1 \frac{dx}{1+x^2}$； (2) $\int_0^1 e^{-x^2}\,dx$； (3) $\int_1^2 \frac{e^{-x}}{x}\,dx$。

16. 利用 $n=4$ 的 Gauss - Laguerre 求积公式计算下列积分：

(1) $\int_0^\infty e^{-x}\sqrt{x}\,dx$； (2) $\int_0^\infty e^{-x^2}\,dx$； (3) $\int_0^\infty \frac{dx}{1+x^2}$。

17. 利用 $n=5$ 的 Gauss - Hermite 求积公式计算下列积分：

(1) $\int_{-\infty}^\infty e^{-x^2}\sqrt{1+x^2}\,dx$； (2) $\int_{-\infty}^\infty e^{-x^2}\cos x\,dx$。

18. 利用 $n=4$ 的 Gauss - Chebyshev 求积公式计算下列积分：

(1) $\int_{-1}^1 \sqrt{\frac{x^2}{1-x^2}}\,dx$；

(2) $a_j = \int_{-1}^1 \frac{e^x T_j(x)}{\sqrt{1-x^2}}\,dx \quad (j = 0,1,2)$。

其中 $T_j(x)$ 是 j 次 Chebyshev 多项式。

19. 试确定 A_1, A_2, x_1, x_2，使下列求积公式成为 Gauss 型求积公式：

(1) $\int_0^1 f(x)\ln\frac{1}{x}\,dx \approx A_1 f(x_1) + A_2 f(x_2)$；

(2) $\int_0^1 \frac{1}{\sqrt{x}} f(x)\,dx \approx A_1 f(x_1) + A_2 f(x_2)$。

20. 证明求积公式

$$\int_{-\infty}^\infty e^{-x^2} f(x)\,dx \approx \frac{\sqrt{\pi}}{6}\left[f\left(-\sqrt{\frac{3}{2}}\right) + 4f(0) + f\left(\sqrt{\frac{3}{2}}\right)\right]$$

具有 5 次代数精度。

21. 选择一种 Gauss 型求积公式计算积分 $\int_1^2 \sin\frac{1}{x}\,dx$。

22. 试根据函数 $u=f(x,y)$ 的数表

y \ u \ x	0	0.5	1	1.5	2
1	10	12	13	15	14
1.2	14	15	17	18	16
1.4	15	16	18	17	15
1.6	13	14	16	15	13

使用复化梯形公式(6.35)计算二重积分 $\iint_D (x,y)\mathrm{d}x\mathrm{d}y$，其中 $D=\{(x,y)\mid 0\leqslant x\leqslant 2, 1\leqslant y\leqslant 1.6\}$。

23. 试用步长 $h=\tau=0.5$ 的复化 Simpson 公式(6.37)计算二重积分 $\iint_D \mathrm{e}^{0.2xy}\mathrm{d}x\mathrm{d}y$，其中

$$D=\{(x,y)\mid 0\leqslant x\leqslant 3,\ 0\leqslant y\leqslant x\}$$

第 7 章 常微分方程初值问题的数值解法

7.1 一般概念

设有常微分方程初值问题

$$\begin{cases} y' = f(t,y), \quad t_0 \leqslant t \leqslant T & (7.1) \\ y(t_0) = y_0 & (7.2) \end{cases}$$

设函数 $f(t,y)$在区域

$$D_0 = \{(t,y) \mid t_0 \leqslant t \leqslant T, |y| < \infty\}$$

内连续且对变量 y 满足 Lipschitz(李普希兹)条件,即存在常数 L,对 D_0 内的任何两点(t,u_1)和(t,u_2),不等式

$$|f(t,u_1) - f(t,u_2)| \leqslant L|u_1 - u_2|$$

成立,因而初值问题(7.1)、(7.2)的解 $y(t)$存在且唯一。为了今后的需要,还设解 $y(t)$在区间$[t_0,T]$上足够光滑,因而设 $f(t,y)$在区域 D_0 内也足够光滑。

如果方程(7.1)是一些特殊的微分方程(例如线性方程、可分离变量方程等),则可通过解析方法求出它的通解,再根据初始条件(7.2)确定通解中的任意常数,就得到初值问题(7.1)、(7.2)的解 $y(t)$的解析表达式。然而在实际问题和科学研究中所遇到的微分方程往往很复杂,很多情况下不可能求出它的解析解。有时候即使能求出解析解,也会由于很难从解析解中计算函数 $y(t)$的值而不实用。例如,容易求出初值问题

$$\begin{cases} y' = 1 - y\cos t, \quad 0 \leqslant t \leqslant T \\ y(0) = 0 \end{cases}$$

的解为

$$y(t) = \mathrm{e}^{-\sin t}\int_0^t \mathrm{e}^{\sin x}\mathrm{d}x$$

但是,对给定的 t,要计算 $y(t)$的值还要用数值积分的方法。

鉴于上述的情况,研究初值问题(7.1)、(7.2)的数值解法就十分必要了。

给定步长 $h>0$,取节点

$$t_n = t_0 + nh \quad (n = 0,1,\cdots,M)$$

其中 $t_M \leqslant T$。要求通过数值计算的方法求出问题(7.1)、(7.2)的解 $y(t)$在各个节点 t_n 处的近似值 $y_n \approx y(t_n)$ $(n=1,2,\cdots,M)$。所用的数值计算方法就称为初值问题(7.1)、(7.2)的数值解法,所求出的近似解 $y_n(n=1,2,\cdots,M)$称为初值问题(7.1)、(7.2)的数值解。

初值问题(7.1)、(7.2)的数值解法的一般形式是

$$F(t_n, y_n, y_{n+1}, \cdots, y_{n+k}, h) = 0 \quad (n = 0,1,\cdots,M-k) \tag{7.3}$$

其中 k 是一正整数,函数 F 与函数 f 有关。方程(7.3)称为关于 $y_0,y_1,\cdots,y_M$ 的差分方程。数值解法的实质是用关于 $y_0,y_1,\cdots,y_M$ 的差分方程(7.3)近似代替原微分方程(7.1),并且从

$y_0,y_1,\cdots,y_{k-1}$ 出发，从差分方程(7.3)中依次逐个解出 $y_k,y_{k+1},\cdots,y_M$，从而得到初值问题(7.1)、(7.2)的数值解。

若 $k=1$，则数值解法(7.3)成为

$$F(t_n,y_n,y_{n+1},h)=0 \quad (n=0,1,\cdots,M-1) \tag{7.4}$$

称数值解法(7.4)为单步法。

若 $k\geqslant 2$，则数值解法(7.3)统称为多步法，或具体称为 k 步法。

若差分方程(7.3)能表示为 y_{n+k} 是 $t_n,y_n,y_{n+1},\cdots,y_{n+k-1},h$ 的显函数，即

$$y_{n+k}=G(t_n,y_n,y_{n+1},\cdots,y_{n+k-1},h) \quad (n=0,1,\cdots,M-k) \tag{7.5}$$

则称数值解法(7.5)为显式方法；否则，称数值解法(7.3)为隐式方法。

定义　设 $y(t)$ 是初值问题(7.1)、(7.2)的解，$y_1,y_2,\cdots,y_M$ 是由数值解法(7.3)解出的初值问题(7.1)、(7.2)的数值解，则称误差

$$\varepsilon_n=y(t_n)-y_n$$

为数值解法(7.3)在节点 t_n 处的整体截断误差。

7.2　显式单步法

7.2.1　显式单步法的一般形式

显式单步法的一般形式是

$$y_{n+1}=y_n+h\varphi(t_n,y_n,h) \quad (n=0,1,\cdots,M-1) \tag{7.6}$$

其中函数 $\varphi(t,y,h)$ 与函数 f 有关，并称为增量函数。函数值 $\varphi(t_n,y_n,h)$ 有明显的几何意义。

把单步法(7.6)在点 t_{n+1} 处的整体截断误差 ε_{n+1} 表达成

$$\begin{aligned}\varepsilon_{n+1}=&\ y(t_{n+1})-y_{n+1}=\\&y(t_{n+1})-y(t_n)-h\varphi(t_n,y(t_n),h)+\\&y(t_n)-y_n+h[\varphi(t_n,y(t_n),h)-\varphi(t_n,y_n,h)]\end{aligned} \tag{7.7}$$

由此产生单步法(7.6)的局部截断误差概念。

定义　设 $y(t)$ 是初值问题(7.1)、(7.2)的解，则称

$$R_{n+1}=y(t_{n+1})-y(t_n)-h\varphi(t_n,y(t_n),h) \tag{7.8}$$

为单步法(7.6)在点 t_{n+1} 处的局部截断误差。

定理 7.1　设增量函数 $\varphi(t,y,h)$ 在区域

$$D=\{(t,y,h)\,|\,t_0\leqslant t\leqslant T,|y|<\infty,0\leqslant h\leqslant h_0\}$$

内对变量 y 满足 Lipschitz 条件，即存在常数 K，使对 D 内任何两点 (t,u_1,h) 和 (t,u_2,h)，不等式

$$|\varphi(t,u_1,h)-\varphi(t,u_2,h)|\leqslant K|u_1-u_2|$$

成立，那么，若单步法(7.6)的局部截断误差 R_{n+1} 与 $h^{p+1}(p\geqslant 1)$ 同阶，即

$$R_{n+1}=O(h^{p+1})$$

则单步法(7.6)的整体截断误差 ε_{n+1} 与 h^p 同阶，即有

$$\varepsilon_{n+1}=O(h^p)$$

证　由式(7.7)，有

$$|\varepsilon_{n+1}|\leqslant|R_{n+1}|+|y(t_n)-y_n|+h|\varphi(t_n,y(t_n),h)-\varphi(t_n,y_n,h)|\leqslant$$

$$|R_{n+1}|+|\varepsilon_n|(1+hK)\leqslant$$
$$|R_{n+1}|+|R_n|(1+hK)+|\varepsilon_{n-1}|(1+hK)^2\leqslant\cdots\leqslant$$
$$\sum_{k=0}^{n}|R_{n+1-k}|(1+hK)^k+|\varepsilon_0|(1+hK)^{n+1}$$

记 $R=\max\limits_{1\leqslant n\leqslant M}|R_n|$，并注意到 $\varepsilon_0=0$，就有

$$|\varepsilon_{n+1}|\leqslant R\sum_{k=0}^{n}(1+hK)^k=\frac{R}{hK}[(1+hK)^{n+1}-1]\leqslant$$
$$\frac{R}{hK}(\mathrm{e}^{(n+1)hK}-1)\leqslant\frac{1}{hK}O(h^{p+1})(\mathrm{e}^{(T-t_0)K}-1)=O(h^p)$$

由此可知，ε_{n+1} 可表示为

$$\varepsilon_{n+1}=c(t_{n+1})h^p+O(h^{p+1})$$

证毕。

由定理 7.1 可知，局部截断误差关于步长 h 的阶的高低可说明单步法(7.6)的精度高低。因此，可以从提高局部截断误差关于步长 h 的阶入手去构造精度较高的数值解法。

定义 若单步法(7.6)的局部截断误差(7.8)与 h^{p+1}(p 为正整数)同阶，即

$$R_{n-1}=O(h^{p+1})$$

则称单步法(7.6)是 p 阶方法。

7.2.2 Runge - Kutta 方法

求解初值问题(7.1)、(7.2)的显式单步法

$$\begin{cases}y_{n+1}=y_n+h\sum\limits_{i=1}^{N}c_ik_i\\ k_1=f(t_n,y_n)\\ k_i=f\left(t_n+a_ih,y_n+h\sum\limits_{j=1}^{i-1}b_{ij}k_j\right)\quad(i=2,3,\cdots,N)\\ a_i=\sum\limits_{j=1}^{i-1}b_{ij}\quad(i=2,3,\cdots,N)\\ \quad(n=0,1,\cdots,M-1)\end{cases}\tag{7.9}$$

称为显式 Runge - Kutta(龙格-库塔)方法，简称 R - K 方法，其中正整数 N 称为 R - K 方法的级，所有 c_i,a_i,b_{ij} 都是待定常数。

根据定义(7.8)，N 级 R - K 方法(7.9)的局部截断误差为

$$R_{n+1}=y(t_{n+1})-y(t_n)-h\sum_{i=1}^{N}c_ik_i\tag{7.10}$$

其中 $k_1,k_2,\cdots,k_N$ 中的 y_n 都换成 $y(t_n)$。如果系数 c_i,a_i,b_{ij} 能使局部截断误差(7.10)与 h^{p+1} 同阶，则相应的 N 级 R - K 方法就是 p 阶方法。一般希望，在 N 确定的情况下，选择一组系数 c_i,a_i,b_{ij}，使阶数 p 达到最高。

一级 R - K 方法的形式为

$$y_{n+1}=y_n+hc_1f(t_n,y_n)\tag{7.11}$$

利用 $y(t_{n+1})$ 在 t_n 处的 Taylor 展开式

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{h^2}{2!}y''(t_n) + \cdots \tag{7.12}$$

可知，方法(7.11)的局部截断误差为

$$R_{n+1} = y(t_{n+1}) - y(t_n) - c_1 hf(t_n, y(t_n)) = y(t_{n+1}) - y(t_n) - c_1 hy'(t_n) =$$

$$(1 - c_1)hy'(t_n) + \frac{h^2}{2!}y''(t_n) + \cdots$$

由此看出，只有当 $c_1 = 1$ 时，方法(7.11)取得最高阶——一阶，所得的一级一阶 R－K 方法为

$$y_{n+1} = y_n + hf(t_n, y_n) \tag{7.13}$$

方法(7.13)又称为 Euler 法，它的局部截断误差为

$$R_{n+1} = \frac{h^2}{2}y''(t_n) + O(h^3)$$

其中$\frac{h^2}{2}y''(t_n)$称为局部截断误差的主项。

二级 R－K 方法的形式为

$$\begin{cases} y_{n+1} = y_n + h(c_1 k_1 + c_2 k_2) \\ k_1 = f(t_n, y_n) \\ k_2 = f(t_n + a_2 h, y_n + a_2 hk_1) \end{cases} \tag{7.14}$$

因 $f(t_n, y(t_n)) = y'(t_n)$，故方法(7.14)的局部截断误差为

$$R_{n+1} = y(t_{n+1}) - y(t_n) - c_1 hy'(t_n) - c_2 hf(t_n + a_2 h, y(t_n) + a_2 hy'(t_n)) \tag{7.15}$$

由二元函数 f 在点$(t_n, y(t_n))$处的 Taylor 展开式，可得

$$f(t_n + a_2 h, y(t_n) + a_2 hy'(t_n)) = f(t_n, y(t_n)) + a_2 hf'_t + a_2 hy'(t_n)f'_y +$$

$$\frac{1}{2!}[a_2^2 h^2 f''_{tt} + 2a_2^2 h^2 y'(t_n)f''_{ty} + a_2^2 h^2 y'^2(t_n)f''_{yy}] + O(h^3) \tag{7.16}$$

其中 f'_t、f'_y、f''_{tt}、f''_{ty}、f''_{yy} 均在点$(t_n, y(t_n))$处取值。把式(7.12)和式(7.16)代入式(7.15)，并注意到

$$y'(t) = f(t, y(t)), \quad y''(t) = f'_t + f'_y y'(t)$$

$$y'''(t) = f''_{tt} + 2f''_{ty}y'(t) + f''_{yy}y'^2(t) + f'_y y''(t)$$

可得

$$R_{n+1} = (1 - c_1 - c_2)hy'(t_n) + \left(\frac{1}{2} - a_2 c_2\right)h^2 y''(t_n) +$$

$$\left(\frac{1}{6} - \frac{1}{2}a_2^2 c_2\right)h^3 y'''(t_n) + \frac{1}{2}a_2^2 c_2 h^3 y''(t_n)f'_y + O(h^4) \tag{7.17}$$

令

$$\begin{cases} 1 - c_1 - c_2 = 0 \\ \frac{1}{2} - a_2 c_2 = 0 \end{cases} \tag{7.18}$$

则 $a_2 \neq 0, c_2 \neq 0$，因而由式(7.17)可看出，二级 R－K 方法(7.14)能达到的最高阶数是二阶，并且凡是满足条件(7.18)的系数 a_2, c_1, c_2 都能使相应的二级 R－K 方法(7.14)成为二阶方法。由式(7.18)可知，二级二阶 R－K 方法的局部截断误差为

$$R_{n+1} = \left(\frac{1}{6} - \frac{a_2}{4}\right)h^3 y'''(t_n) + \frac{a_2}{4}h^3 y''(t_n)f'_y + O(h^4)$$

式(7.18)是三个未知数、二个方程的方程组,可以有多组不同的解,将其解代回式(7.14)就构成不同的二级二阶 R-K 方法。

下面是常用的三种二级二阶 R-K 方法。

取 $c_1=c_2=\frac{1}{2},a_2=1$,式(7.14)成为

$$\begin{cases} y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2) \\ k_1 = f(t_n, y_n) \\ k_2 = f(t_n + h, y_n + hk_1) \end{cases} \tag{7.19}$$

方法(7.19)又称为改进的 Euler 法。

取 $c_1=0,c_2=1,a_2=\frac{1}{2}$,则式(7.14)成为

$$\begin{cases} y_{n+1} = y_n + hk_2 \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right) \end{cases} \tag{7.20}$$

式(7.20)又称为中点公式。

取 $c_1=\frac{1}{4},c_2=\frac{3}{4},a_2=\frac{2}{3}$,则式(7.14)成为

$$\begin{cases} y_{n+1} = y_n + \frac{h}{4}(k_1 + 3k_2) \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_1\right) \end{cases} \tag{7.21}$$

式(7.21)又称为 Heun(休恩)方法。

三级 R-K 方法的形式是

$$\begin{cases} y_{n+1} = y_n + h(c_1k_1 + c_2k_2 + c_3k_3) \\ k_1 = f(t_n, y_n) \\ k_2 = f(t_n + a_2h, y_n + hb_{21}k_1) \\ k_3 = f(t_n + a_3h, y_n + h(b_{31}k_1 + b_{32}k_2)) \end{cases} \tag{7.22}$$

完全仿照前述的方法可推出三级 R-K 方法能达到的最高阶是三阶,并且凡是满足条件

$$\begin{cases} c_1 + c_2 + c_3 = 1 \\ c_2a_2 + c_3a_3 = \frac{1}{2} \\ c_2a_2^2 + c_3a_3^2 = \frac{1}{3} \\ c_3a_2b_{32} = \frac{1}{6} \\ a_2 = b_{21} \\ a_3 = b_{31} + b_{32} \end{cases} \tag{7.23}$$

的系数 $c_1, c_2, c_3, a_2, b_{21}, a_3, b_{31}, b_{32}$ 都能使相应的三级 R-K 方法(7.22)成为三阶方法。方程组(7.23)含八个未知数、六个方程。可有多组不同的解,因而有多种三级三阶 R-K 方法。

下面是两种具体的三级三阶 R-K 方法。

Heun 三阶方法:

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{4}(k_1 + 3k_3) \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \dfrac{1}{3}h, y_n + \dfrac{1}{3}hk_1\right) \\ k_3 = f\left(t_n + \dfrac{2}{3}h, y_n + \dfrac{2}{3}hk_2\right) \end{cases} \tag{7.24}$$

Kutta 三阶方法:

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{6}(k_1 + 4k_2 + k_3) \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \dfrac{1}{2}h, y_n + \dfrac{1}{2}hk_1\right) \\ k_3 = f(t_n + h, y_n - hk_1 + 2hk_2) \end{cases} \tag{7.25}$$

四级 R-K 方法的形式为

$$\begin{cases} y_{n+1} = y_n + h(c_1k_1 + c_2k_2 + c_3k_3 + c_4k_4) \\ k_1 = f(t_n, y_n) \\ k_2 = f(t_n + a_2h, y_n + hb_{21}k_1) \\ k_3 = f(t_n + a_3h, y_n + h(b_{31}k_1 + b_{32}k_2)) \\ k_4 = f(t_n + a_4h, y_n + h(b_{41}k_1 + b_{42}k_2 + b_{43}k_3)) \end{cases} \tag{7.26}$$

经过类似上述的方法可以证明,四级 R-K 方法(7.26)能达到的最高阶是四阶,并且可以构造多种四级四阶 R-K 方法,其中最经常使用的是经典 R-K 方法:

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \dfrac{1}{2}h, y_n + \dfrac{1}{2}hk_1\right) \\ k_3 = f\left(t_n + \dfrac{1}{2}h, y_n + \dfrac{1}{2}hk_2\right) \\ k_4 = f(t_n + h, y_n + hk_3) \end{cases} \tag{7.27}$$

下面的四级四阶 R-K 方法被称为 Gill(基尔)方法:

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{6}\left[k_1 + (2-\sqrt{2})k_2 + (2+\sqrt{2})k_3 + k_4\right] \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + \dfrac{1}{2}h, y_n + \dfrac{1}{2}hk_1\right) \\ k_3 = f\left(t_n + \dfrac{1}{2}h, y_n + \dfrac{\sqrt{2}-1}{2}hk_1 + \left(1-\dfrac{\sqrt{2}}{2}\right)hk_2\right) \\ k_4 = f\left(t_n + h, y_n - \dfrac{\sqrt{2}}{2}hk_2 + \left(1+\dfrac{\sqrt{2}}{2}\right)hk_3\right) \end{cases} \tag{7.28}$$

前面已看到，N 级 R－K 方法在 $N=1,2,3,4$ 时，可分别得到最高阶数一、二、三、四阶，但是，通常 N 级 R－K 方法的最高阶不一定是 N 阶。设 $p(N)$是 N 级 R－K 方法可达到的最高阶数，可以证明：

$$p(5)=4,\quad p(6)=5,\quad p(7)=6,\quad p(8)=6,\quad p(9)=7$$

例 1 分别用 Euler 法(7.13)、改进的 Euler 法(7.19)和经典 R－K 法(7.27)求解初值问题

$$\begin{cases} y' = 1 - \dfrac{2ty}{1+t^2}, \quad 0 \leqslant t \leqslant 2 \\ y(0) = 0 \end{cases}$$

取步长 $h=0.5$，并与精确解 $y(t)=\dfrac{t(3+t^2)}{3(1+t^2)}$作比较。

解 用 Euler 法求解的计算公式为

$$\begin{cases} y_0 = 0 \\ y_{n+1} = y_n + h\left(1 - \dfrac{2t_ny_n}{1+t_n^2}\right) \quad (n=0,1,2,3) \end{cases}$$

其中 $h=0.5$，$t_n=0.5n$。计算结果见表 7－1。

表 7－1 Euler 法求解结果与精确解比较

n	t_n	y_n	$y(t_n)$	$y(t_n)-y_n$
0	0	0	0	0
1	0.5	0.500 000	0.433 333	0.066 667
2	1.0	0.800 000	0.666 667	0.133 333
3	1.5	0.900 000	0.807 692	0.092 308
4	2.0	0.984 615	0.933 333	0.051 282

用改进的 Euler 法求解，计算公式为

$$\begin{cases} y_0 = 0 \\ k_1 = 1 - \dfrac{2t_ny_n}{1+t_n^2} \\ k_2 = 1 - \dfrac{2(t_n+h)(y_n+hk_1)}{1+(t_n+h)^2} \\ y_{n+1} = y_n + \dfrac{h}{2}(k_1+k_2) \quad (n=0,1,2,3) \end{cases}$$

其中 $h=0.5, t_n=0.5n$。计算结果见表 7-2。

表 7-2　改进的 Euler 法求解结果与精确解比较

n	t_n	y_n	k_1	k_2	$y(t_n)-y_n$
0	0	0	1.000 000	0.600 000	0
1	0.5	0.400 000	0.680 000	0.260 000	0.033 333
2	1.0	0.635 000	0.365 000	0.245 385	0.031 667
3	1.5	0.787 596	0.272 988	0.260 728	0.020 096
4	2.0	0.921 025			0.012 308

用经典 R-K 法(7.27)求解，计算公式为

$$\begin{cases} y_0 = 0 \\ k_1 = 1 - \dfrac{2t_n y_n}{1+t_n^2} \\ k_2 = 1 - \dfrac{2\left(t_n + \frac{1}{2}h\right)\left(y_n + \frac{1}{2}hk_1\right)}{1+\left(t_n+\frac{1}{2}h\right)^2} \\ k_3 = 1 - \dfrac{2\left(t_n + \frac{1}{2}h\right)\left(y_n + \frac{1}{2}hk_2\right)}{1+\left(t_n+\frac{1}{2}h\right)^2} \\ k_4 = 1 - \dfrac{2(t_n+h)(y_n+hk_3)}{1+(t_n+h)^2} \\ y_{n+1} = y_n + \dfrac{h}{6}(k_1+2k_2+2k_3+k_4) \\ \qquad (n=0,1,2,3) \end{cases}$$

其中 $h=0.5, t_n=0.5n$。计算结果见表 7-3。

表 7-3　经典 R-K 法求解结果与精确解比较

n	t_n	y_n	k_1	k_2	k_3	k_4	$y(t_n)-y_n$
0	0	0	1.000 000	0.882 353	0.896 194	0.641 522	0
1	0.5	0.433 218	0.653 426	0.427 289	0.481 561	0.326 001	0.000 115
2	1.0	0.666 312	0.333 688	0.268 552	0.284 439	0.253 663	0.000 355
3	1.5	0.807 423	0.254 686	0.249 518	0.250 632	0.253 809	0.000 269
4	2.0	0.933 156					0.000 177

7.2.3　相容性、收敛性和绝对稳定性

在单步法(7.6)中，如要把 y_n 和 y_{n+1} 分别换成初值问题(7.1)、(7.2)的解 $y(t)$ 和 $y(t+h)$，则得到关于 $y(t)$ 的一个近似方程

$$\frac{y(t+h)-y(t)}{h} \approx \varphi(t, y(t), h) \tag{7.29}$$

差分方程(7.6)的解 $y_1, y_2, \cdots, y_M$ 是否可作为初值问题(7.1)、(7.2)的近似解，应取决于当

$h \to 0$时近似方程(7.29)的极限状态能否成为微分方程(7.1)。由于

$$\lim_{h\to 0}\frac{y(t+h)-y(t)}{h}=y'(t)$$

因此,要使近似方程(7.29)的极限状态成为微分方程(7.1),须且只须极限

$$\lim_{h\to 0}\varphi(t,y(t),h)=f(t,y(t)) \tag{7.30}$$

成立。总假定函数 $\varphi(t,y(t),h)$是连续函数,因而极限(7.30)可表示为

$$\varphi(t,y(t),0)=f(t,y(t)) \tag{7.31}$$

定义 单步法(7.6)称为与微分方程(7.1)相容,如果条件(7.31)成立;并称条件(7.31)为相容性条件。

利用相容性条件(7.31)容易验证,Euler 法(7.13)、系数满足方程(7.18)的二级 R-K 方法(7.14)、系数满足方程(7.23)的三级 R-K 方法(7.22)以及四级 R-K 方法(7.27)和(7.28)都与微分方程(7.1)相容。

定理 7.2 设增量函数 $\varphi(t,y,h)$在区域 $D=\{(t,y,h)\mid t_0\leqslant t\leqslant T,|y|<\infty,0\leqslant h\leqslant h_0\}$上连续,且对变量 h 满足 Lipschitz 条件,则单步法(7.6)与微分方程(7.1)相容的充分必要条件是单步法(7.6)至少是一阶的方法。

证 必要性

$$\begin{aligned}R_{n+1}&=y(t_{n+1})-y(t_n)-h\varphi(t_n,y(t_n),h)=\\&[y(t_n)+hy'(t_n)+O(h^2)]-y(t_n)-\\&h[\varphi(t_n,y(t_n),h)-\varphi(t_n,y(t_n),0)]-h\,f(t_n,\ y(t_n))\end{aligned}$$

$$|R_{n+1}|\leqslant O(h^2)+L_1h^2=O(h^2)$$

式中 L_1 是 Lipschitz 常数。

充分性 由

$$R_{n+1}=y(t_{n+1})-y(t_n)-h\varphi(t_n,y(t_n),h)=O(h^{p+1})$$

其中 $p\geqslant 1$,可得

$$\frac{y(t_{n+1})-y(t_n)}{h}=\varphi(t_n,y(t_n),h)+O(h^p)$$

令 $h\to 0$,由上式得

$$y'(t_n)=\varphi(t_n,y(t_n),0)$$

因而等式

$$\varphi(t_n,y(t_n),0)=f(t_n,y(t_n))$$

成立。

证毕。

定义 求解初值问题(7.1)、(7.2)的单步法(7.6)叫做收敛的,如果对任意的 y_0 及任意的 $t\in(t_0,T)$,极限

$$\lim_{\substack{h\to 0(n\to\infty)\\ t_n=t}}y_n=y(t)$$

成立,其中 $y(t)$是初值问题(7.1)、(7.2)的解。

由这个定义可知,在区间(t_0,T)内任取一固定点 t 作为节点 t_n,从 y_0 开始,使用单步法(7.6)以不同的步长 $h=\frac{t_n-t_0}{n}$计算相应的 y_n 值,如果 $h\to 0(n\to\infty)$时,y_n 收敛于初值问题

(7.1)、(7.2)的解 $y(t)$，则称单步法(7.6)是收敛的。因此，如果单步法(7.6)在区间(t_0,T)的任一点 t_n 处的整体截断误差 ε_n 满足

$$\lim_{h\to 0(n\to\infty)} \varepsilon_n = 0$$

则单步法(7.6)就是收敛的；反之也成立。

定理 7.3　设增量函数 $\varphi(t,y,h)$在区域

$$D = \{(t,y,h)\,|\,t_0 \leqslant t \leqslant T, |y| < \infty, 0 \leqslant h \leqslant h_0\}$$

上连续，并对变量 y 和 h 满足 Lipschitz 条件。如果单步法(7.6)与微分方程(7.1)相容，则单步法(7.6)是收敛的。

证　根据定理 7.2 和定理 7.1 即可知此定理成立。

证毕。

可以证明，只要微分方程(7.1)的右端函数 $f(t,y)$在区域 D_0 内连续和有界，且对变量 t 和 y 满足 Lipschitz 条件，那么，Euler 法(7.13)、系数满足方程(7.18)的二级 R-K 方法(7.14)、系数满足方程(7.23)的三级 R-K 方法(7.22)以及四级 R-K 方法(7.27)和(7.28)，这些方法的增量函数 $\varphi(t,y,h)$就都在区域 D 内连续且对变量 y 和 h 满足 Lipschitz 条件。又因为上述这些方法的增量函数都满足相容性条件(7.31)，因而根据定理 7.3，用上述这些方法求解初值问题(7.1)、(7.2)时都是收敛的。

关于单步法收敛的概念和收敛定理都是在计算过程无任何舍入误差的前提下建立起来的。整体截断误差 $\varepsilon_n = y(t_n) - y_n$ 中的 y_n 是以 y_0 为初始值由单步法(7.6)经过精确计算得到的。但是，实际计算时通常都会有舍入误差。特别是式(7.6)是一个递推算式，凡是递推算式都要考虑舍入误差的积累是否会得到控制，也就是要考虑数值稳定性的问题。一个单步法(7.6)，即使它已满足相容性条件，并且又是收敛的，然而在用它计算 $y_1, y_2, \cdots$时，如果舍入误差的积累越来越大，那么由它算出的 $y_1, y_2, \cdots$仍然不能作为初值问题(7.1)、(7.2)的近似解使用。因此必须讨论单步法的数值稳定性问题。

对于给定的微分方程(7.1)和给定的步长 h，如果单步法(7.6)在计算 y_n 时有大小为 δ 的误差，但由此引起 $y_m(m>n)$的误差按绝对值又都不超过 δ，则称单步法(7.6)是绝对稳定的。然而，此定义太依赖于微分方程本身。为摆脱这种依赖性，在定义方法的绝对稳定性时，都针对同一类型微分方程，就是线性常系数微分方程

$$y' = \lambda y \tag{7.32}$$

其中 λ 为复常数，并称式(7.32)为模型方程。

定义　设步长为 $h>0$ 的单步法(7.6)用于求解模型方程(7.32)的初值问题，又设初值 y_0 有误差 e_0，如果在计算后面的 y_n 时，由 e_0 所引起的误差 e_n 满足

$$\lim_{n\to\infty} e_n = 0$$

则称单步法(7.6)对于所用的步长 h 和复数 λ 是绝对稳定的。

从定义可知，单步法(7.6)是否绝对稳定，与模型方程中的 λ 以及所用的步长 h 有关。如果对 $h\lambda$ 复平面上的一个区域 G，当 $h\lambda \in G$ 时，都使单步法(7.6)绝对稳定，则称 G 为单步法(7.6)的绝对稳定区域。

用 Euler 法(7.13)求解模型方程(7.32)的初值问题，计算公式为

$$y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda) y_n$$

设 y_n 有误差 e_n，$e_n = y_n - \tilde{y}_n$，则有

$$\tilde{y}_{n+1} = (1 + h\lambda)\tilde{y}_n$$

$$e_{n+1} = (1 + h\lambda)e_n = \cdots = (1 + h\lambda)^{n+1} e_0$$

当且仅当 $|1+h\lambda|<1$ 时有 $\lim\limits_{n\to\infty} e_n = 0$。所以，Euler 法(7.13)的绝对稳定区域为

$$|1 + h\lambda| < 1$$

当 λ 为实数时，得 Euler 法(7.13)的绝对稳定区间 $-2<h\lambda<0$。

用二级二阶 R－K 方法求解模型方程(7.32)的初值问题，计算公式为

$$\begin{aligned} y_{n+1} &= y_n + h[c_1\lambda y_n + c_2\lambda(y_n + a_2 h\lambda y_n)] = \\ &[1 + (c_1 + c_2)h\lambda + c_2 a_2 h^2\lambda^2]y_n = \\ &\left(1 + h\lambda + \frac{1}{2}h^2\lambda^2\right)y_n \end{aligned}$$

由此可知，二级二阶 R－K 方法的绝对稳定区域是

$$\left|1 + h\lambda + \frac{(h\lambda)^2}{2}\right| < 1$$

当 λ 为实数时，得绝对稳定区间 $-2<h\lambda<0$。

同理可推出三级三阶 R－K 方法的绝对稳定区域

$$\left|1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{6}\right| < 1$$

当 λ 为实数时，得绝对稳定区间 $-2.51<h\lambda<0$。

同理又可推出四级四阶 R－K 方法的绝对稳定区域

$$\left|1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{6} + \frac{(h\lambda)^4}{24}\right| < 1$$

当 λ 为实数时，得绝对稳定区间 $-2.78<h\lambda<0$。

上述四种方法的绝对稳定区域见图 7－1，图中的 N 表示 R－K 方法的级。从图 7－1 看

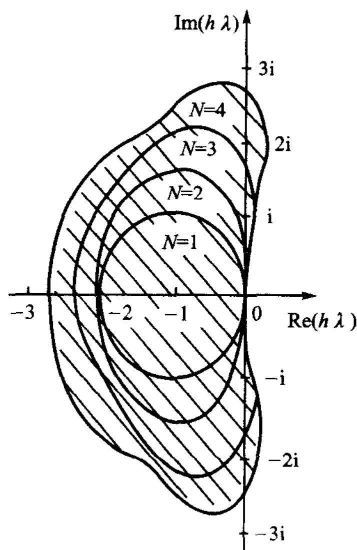


图 7－1 R－K 方法的绝对稳定区域

出，N 级 R－K 方法的绝对稳定区域随着 N 的增大而扩大。

绝对稳定区域越大，那么，为使数值解法的计算过程具有数值稳定性，对步长 h 的限制就越小。

例 2　用 Euler 法求解

$$\begin{cases} y' = -5y + t, & t_0 \leqslant t \leqslant T \\ y(t_0) = y_0, \end{cases}$$

从绝对稳定性考虑，对步长 h 有何限制？

解　这里 $\lambda = -5$，由

$$|1 + h\lambda| = |1 - 5h| < 1$$

得到对 h 的限制为 $0 < h < 0.4$。

对于一般微分方程(7.1)，对其右端函数 $f(t,y)$ 进行线性化处理后可知 $\lambda = \frac{\partial f}{\partial y}$。这时 λ 将是变化的，但只要 $h\lambda = h\frac{\partial f}{\partial y}$ 属于所用方法的绝对稳定区域，则该方法就是绝对稳定的。

例 3　用经典 R－K 方法(7.27)求解

$$\begin{cases} y' = 1 - \dfrac{10ty}{1+t^2}, & 0 \leqslant t \leqslant 10 \\ y(0) = 0 \end{cases}$$

从绝对稳定性考虑，对步长 h 有何限制？

解　对于所给微分方程，有

$$\lambda = \frac{\partial f}{\partial y} = -\frac{10t}{1+t^2} \leqslant 0, \quad 0 \leqslant t \leqslant 10$$

$$\max_{0\leqslant t\leqslant 10} |\lambda| = \max_{0\leqslant t\leqslant 10} \frac{10t}{1+t^2} = 5$$

根据经典 R－K 方法的绝对稳定区间 $-2.78 < h\lambda < 0$，步长 h 应满足

$$-2.78 < -5h < 0, \quad 0 < h < 0.556$$

在使用任何数值解法时，步长 h 若不满足绝对稳定性的要求，由于计算机的位数有限，可能会产生很大的误差，计算结果会完全失真，试看下述实例。

用经典 R－K 方法(7.27)求解

$$y' = -20y, \quad y(0) = 1$$

步长分别取 0.1 和 0.2。当 $h = 0.1$ 时，$-20h$ 属于绝对稳定区间 $(-2.78, 0)$；当 $h = 0.2$ 时，$-20h$ 不属于绝对稳定区间。在 10 位计算器上计算，计算结果 y_n 的全误差 $y(t_n) - y_n$ 见表 7－4。

表 7－4　两种步长的数值稳定性比较

t	$y(t_n) - y_n$　$(h=0.1)$	$y(t_n) - y_n$　$(h=0.2)$
0.0	0	0
0.2	−0.092 795	−4.98
0.4	−0.012 010	−25.0
0.6	−0.001 366	−125.0
0.8	−0.000 152	−625.0
1.0	−0.000 017	−3125.0

7.3 线性多步法

7.3.1 线性多步法的一般形式

求解初值问题(7.1)、(7.2)的线性 k 步法的一般形式为

$$\sum_{j=0}^{k}\alpha_j y_{n+j} = h\sum_{j=0}^{k}\beta_j f_{n+j} \quad (n = 0,1,\cdots,M-k) \tag{7.33}$$

其中,$\alpha_j,\beta_j(j=0,1,\cdots,k)$是常系数,$\alpha_k=1$,$|\alpha_0|+|\beta_0|\neq 0$,$f_{n+j}=f(t_{n+j},y_{n+j})$。式(7.33)的左右两边分别是关于 $y_n,y_{n+1},\cdots,y_{n+k}$ 和关于 $f_n,f_{n+1},\cdots,f_{n+k}$ 的线性表达式,并且其中必含有 y_n 和 y_{n+k},故称式(7.33)为线性 k 步法。若 $\beta_k=0$,则式(7.33)是显式方法;若 $\beta_k\neq 0$,则式(7.33)就是隐式方法。

定义 设 $y(t)$是初值问题(7.1)、(7.2)的解,则称

$$R_{n+k} = \sum_{j=0}^{k}\alpha_j y(t_{n+j}) - h\sum_{j=0}^{k}\beta_j f(t_{n+j},y(t_{n+j}))$$

为线性 k 步法(7.33)在点 t_{n+k}处的局部截断误差。若

$$R_{n+k} = O(h^{p+1}) \quad (p\text{ 为正整数})$$

则称线性 k 步法(7.33)是 p 阶方法。

显然,k 步法(7.33)的阶数 p 与 k 步法(7.33)中的系数 $\alpha_j,\beta_j(j=0,1,\cdots,k)$有关。利用 Taylor 级数把局部截断误差 R_{n+k}表示成关于步长 h 的幂级数就可找出它们之间的关系。

注意到

$$f(t_{n+j},y(t_{n+j})) = y'(t_{n+j}), \quad t_{n+j} = t_n + jh$$

则有

$$R_{n+k} = \sum_{j=0}^{k}\alpha_j y(t_n+jh) - h\sum_{j=0}^{k}\beta_j y'(t_n+jh) \tag{7.34}$$

将 $y(t_n+jh)$和 $y'(t_n+jh)$在点 t_n 处展成 Taylor 级数

$$y(t_n+jh) = y(t_n) + jhy'(t_n) + \sum_{r=2}^{\infty}\frac{1}{r!}(jh)^r y^{(r)}(t_n)$$

$$y'(t_n+jh) = y'(t_n) + \sum_{r=2}^{\infty}\frac{1}{(r-1)!}(jh)^{r-1}y^{(r)}(t_n)$$

代入式(7.34)得

$$R_{n+k} = c_0 y(t_n) + c_1 hy'(t_n) + c_2h^2y''(t_n) + \cdots + c_rh^ry^{(r)}(t_n) + \cdots$$

其中

$$\begin{cases} c_0 = \alpha_0 + \alpha_1 + \cdots + \alpha_k \\ c_1 = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k - (\beta_0 + \beta_1 + \cdots + \beta_k) \\ \qquad\vdots \\ c_r = \sum\limits_{j=1}^{k}\dfrac{j^r\alpha_j}{r!} - \sum\limits_{j=1}^{k}\dfrac{j^{r-1}\beta_j}{(r-1)!} \quad (r = 2,3,\cdots) \end{cases} \tag{7.35}$$

如果有一组系数 $\alpha_j,\beta_j(j=0,1,\cdots,k)$使得

$$c_0 = c_1 = \cdots = c_p = 0, \quad c_{p+1} \neq 0$$

则有

$$R_{n+k} = c_{p+1} h^{p+1} y^{(p+1)}(t_n) + O(h^{p+2})$$

此时，k 步法(7.33)是 p 阶方法。$c_{p+1} h^{p+1} y^{(p+1)}(t_n)$ 称为局部截断误差的主项，c_{p+1} 称为误差常数。上述构造线性多步法的方法称为待定系数法，又叫 Taylor 级数法。

例 4　试确定二步法

$$y_{n+2} = -\alpha_1 y_{n+1} + h(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n)$$

中的系数，使它的阶尽可能高，并求它的局部截断误差和阶数。

解　已知 $\alpha_2 = 1, \alpha_0 = 0$，令公式(7.35)中的 $c_0 = c_1 = c_2 = c_3 = 0$，得方程组

$$\begin{cases} \alpha_1 + 1 = 0 \\ \alpha_1 + 2 - (\beta_0 + \beta_1 + \beta_2) = 0 \\ \dfrac{1}{2}(\alpha_1 + 4) - (\beta_1 + 2\beta_2) = 0 \\ \dfrac{1}{6}(\alpha_1 + 8) - \dfrac{1}{2}(\beta_1 + 4\beta_2) = 0 \end{cases}$$

解得 $\alpha_1 = -1, \beta_0 = -\dfrac{1}{12}, \beta_1 = \dfrac{8}{12}, \beta_2 = \dfrac{5}{12}$，得到二步法

$$y_{n+2} = y_{n+1} + \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n) \tag{7.36}$$

此时公式(7.35)中的 c_4 为

$$c_4 = \frac{1}{24}(\alpha_1 + 16) - \frac{1}{6}(\beta_1 + 8\beta_2) = -\frac{1}{24}$$

故二步法(7.36)的局部截断误差为

$$R_{n+2} = -\frac{1}{24} h^4 y^{(4)}(t_n) + O(h^5)$$

并可知二步法(7.36)是三阶方法，它又称为二步隐式 Adams(亚当斯)方法。

用类似的计算过程，可获得下面一些常用的线性多步法。

三步显式 Adams 方法(三阶)：

$$y_{n+3} = y_{n+2} + \frac{h}{12}(23f_{n+2} - 16f_{n+1} + 5f_n) \tag{7.37}$$

其局部截断误差为

$$R_{n+3} = \frac{3}{8} h^4 y^{(4)}(t_n) + O(h^5)$$

三步隐式 Adams 方法(四阶)：

$$y_{n+3} = y_{n+2} + \frac{h}{24}(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n) \tag{7.38}$$

其局部截断误差为

$$R_{n+3} = -\frac{19}{720} h^5 y^{(5)}(t_n) + O(h^6)$$

四步显式 Adams 方法(四阶)：

$$y_{n+4} = y_{n+3} + \frac{h}{24}(55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n) \tag{7.39}$$

其局部截断误差为

$$R_{n+4} = \frac{251}{720}h^5 y^{(5)}(t_n) + O(h^6)$$

显式 Milne(密伦)公式(四步四阶)：

$$y_{n+4} = y_n + \frac{4}{3}h(2f_{n+3} - f_{n+2} + 2f_{n+1}) \tag{7.40}$$

其局部截断误差为

$$R_{n+4} = \frac{14}{45}h^5 y^{(5)}(t_n) + O(h^6)$$

Hamming(哈明)公式(三步四阶)：

$$y_{n+3} = \frac{1}{8}(9y_{n+2} - y_n) + \frac{3h}{8}(f_{n+3} + 2f_{n+2} - f_{n+1}) \tag{7.41}$$

其局部截断误差为

$$R_{n+3} = -\frac{1}{40}h^5 y^{(5)}(t_n) + O(h^6)$$

当 $k=1$ 时，式(7.33)成为线性单步法

$$y_{n+1} + \alpha_0 y_n = h(\beta_0 f_n + \beta_1 f_{n+1})$$

线性单步法除了 Euler 法(7.13)外，还有以下两种。

梯形法(二阶隐式方法)：

$$y_{n+1} = y_n + \frac{h}{2}(f_n + f_{n+1}) \tag{7.42}$$

其局部截断误差为

$$R_{n+1} = -\frac{1}{12}h^3 y'''(t_n) + O(h^4)$$

向后 Euler 法(一阶隐式方法)：

$$y_{n-1} = y_n + hf_{n+1} \tag{7.43}$$

其局部截断误差为

$$R_{n+1} = -\frac{1}{2}h^2 y''(t_n) + O(h^3)$$

形如

$$\sum_{j=0}^{k}\alpha_j y_{n+j} = h\beta_k f_{n+k} \tag{7.44}$$

的线性 k 步法称为 Gear(吉尔)方法。Gear 按照式(7.44)的形式对于 k 从 1 到 6 构造了 k 阶 k 步法。k 从 1 到 6 的 Gear 方法的系数见表 7-5。

例 5 构造二阶二步 Gear 方法。

解 此时，式(7.44)中 $k=2$。取 $\alpha_2=1$，由公式(7.36)，令

$$\begin{cases} c_0 = \alpha_0 + \alpha_1 + 1 = 0 \\ c_1 = \alpha_1 + 2 - \beta_2 = 0 \\ c_2 = \dfrac{1}{2}(\alpha_1 + 4) - 2\beta_2 = 0 \end{cases}$$

解得 $\alpha_1 = -\frac{4}{3}$，$\alpha_0 = \frac{1}{3}$，$\beta_2 = \frac{2}{3}$。因

$$c_3 = \frac{1}{3!}(\alpha_1 + 8\alpha_2) - \frac{1}{2}(4\beta_2) = -\frac{2}{9} \neq 0$$

故得二阶二步 Gear 方法为

$$y_{n+2} - \frac{4}{3}y_{n+1} + \frac{1}{3}y_n = \frac{2}{3}hf_{n+2}$$

构造线性多步法的方法除了本节所阐述的待定系数法之外，还可以用数值积分法来构造线性多步法。但是，本书不介绍这个方法。

表 7-5　Gear 方法系数表

k	α_6	α_5	α_4	α_3	α_2	α_1	α_0	β_k
1						1	−1	1
2					1	$-\frac{4}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
3				1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$	$\frac{6}{11}$
4			1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	$\frac{12}{25}$
5		1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	$\frac{60}{137}$
6	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{255}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$	$\frac{60}{147}$

使用线性 k 步法求解初值问题(7.1)、(7.2)须要有 k 个开始值 $y_0, y_1, \cdots, y_{k-1}$。除了 y_0 是给定的之外，其余 $k-1$ 个开始值要用单步法计算。所用的单步法应至少与该线性 k 步法同阶，例如可用四阶 R-K 方法计算四阶 k 步法的开始值 $y_1, y_2, \cdots, y_{k-1}$。

7.3.2　预报-校正格式

对于隐式线性 k 步法，一般情况下是一个关于 y_{n+k} 的非线性方程，从中解出 y_{n+k} 一般要用迭代法，迭代的初值最好用同阶的显式方法。在实际计算工作中，往往只迭代一次，因而由此产生了预报-校正格式。用显式方法计算 y_{n+k} 的初值 $\tilde{y}_{n+k}$，称为预报值。把 $\tilde{y}_{n+k}$ 代入隐式方法的右边，计算出新的 y_{n+k}，称为校正值。

常用的且最简单的预报-校正格式有下面几种。

Euler 法与梯形法构成的预报-校正格式：

$$\begin{cases} \tilde{y}_{n+1} = y_n + hf_n \\ y_{n+1} = y_n + \dfrac{h}{2}[f_n + f(t_{n+1}, \tilde{y}_{n+1})] \\ \qquad (n = 0,1,\cdots,M-1) \end{cases}$$

这个格式实际上是改进的 Euler 法(7.19)。

显式 Milne 方法与 Hamming 方法构成的预报-校正格式：

$$\begin{cases} \tilde{y}_{n+1} = y_{n-3} + \dfrac{4}{3}h(2f_{n-2} - f_{n-1} + 2f_n) \\ y_{n+1} = \dfrac{1}{8}(9y_n - y_{n-2}) + \dfrac{3h}{8}[f(t_{n+1}, \tilde{y}_{n+1}) + 2f_n - f_{n-1}] \end{cases}$$

$$(n=3,4,\cdots,M-1)$$

四步显式与三步隐式的 Adams 方法构成的预报-校正格式：

$$\begin{cases}\tilde{y}_{n+1}=y_n+\dfrac{h}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3})\\ y_{n+1}=y_n+\dfrac{h}{24}[9f(t_{n+1},\tilde{y}_{n+1})+19f_n-5f_{n-1}+f_{n-2}]\\ \qquad(n=3,4,\cdots,M-1)\end{cases}$$

7.3.3 相容性和收敛性

在 k 步法(7.33)中，如果把 y_{n+j} 换成初值问题(7.1)、(7.2)的解 $y(t+jh)$，则得到关于 $y(t)$ 的一个近似方程

$$\frac{1}{h}\sum_{j=0}^{k}\alpha_j y(t+jh)\approx\sum_{j=0}^{k}\beta_j f(t+jh,y(t+jh)) \tag{7.45}$$

只有当 $h\to0$ 时近似方程(7.45)的极限状态能成为微分方程(7.1)，才能用差分方程(7.33)的解作为初值问题(7.1)、(7.2)的近似解。由于

$$\frac{1}{h}\sum_{j=0}^{k}\alpha_j y(t+jh)=\frac{y(t)}{h}\sum_{j=0}^{k}\alpha_j+\sum_{j=1}^{k}j\alpha_j\,\frac{y(t+jh)-y(t)}{jh}$$

$$\lim_{h\to0}\sum_{j=1}^{k}j\alpha_j\,\frac{y(t+jh)-y(t)}{jh}=y'(t)\sum_{j=1}^{k}j\alpha_j$$

$$\lim_{h\to0}\sum_{j=0}^{k}\beta_j f(t+jh,y(t+jh))=f(t,y(t))\sum_{j=0}^{k}\beta_j$$

所以，当且仅当

$$\sum_{j=0}^{k}\alpha_j=0,\quad\sum_{j=1}^{k}j\alpha_j=\sum_{j=0}^{k}\beta_j \tag{7.46}$$

成立时，近似方程(7.45)的极限状态($h\to0$)成为微分方程(7.1)。记

$$\rho(\xi)=\sum_{j=0}^{k}\alpha_j\xi^j,\quad\sigma(\xi)=\sum_{j=0}^{k}\beta_j\xi^j$$

则条件(7.46)等价于

$$\rho(1)=0,\quad\rho'(1)=\sigma(1) \tag{7.47}$$

定义 线性 k 步法(7.33)叫做与微分方程(7.1)相容，如果条件(7.47)成立；并称条件(7.47)为线性 k 步法(7.33)的相容性条件。

易知，相容性条件(7.47)相当于公式(7.35)中的 $c_0=0$ 和 $c_1=0$。所以，线性 k 步法(7.33)与微分方程(7.1)相容的充分必要条件是方法(7.33)至少是一阶的方法。同时可看出，方法(7.33)的阶数 p 越高，用差分方程(7.33)逼近微分方程(7.1)的精度就越高。

定义 求解初值问题(7.1)、(7.2)的线性 k 步法(7.33)叫做收敛的，如果有

(1) 由其他单步法计算的 $k-1$ 个开始值 y_j，有

$$\lim_{h\to0}y_j=y_0\quad(j=1,2,\cdots,k-1)$$

(2) 对任意的 $t\in(t_0,T)$，由式(7.33)计算的 y_m 均满足

$$\lim_{\substack{h\to0(m\to\infty)\\ t_m=t}}y_m=y(t)$$

其中 $y(t)$ 是初值问题(7.1)、(7.2)的解。

定义　如果多项式 $\rho(\xi)=\sum_{j=0}^{k}\alpha_j\xi^j$ 的零点的模不大于1，并且模为1的零点都是单零点，则称线性 k 步法(7.33)满足根条件。

定理 7.4　设线性 k 步法(7.33)满足相容性条件(7.47)和根条件，则当计算开始值的单步法收敛时，k 步法(7.33)也是收敛的；此外，若 k 步法(7.33)是 p 阶的，并且开始所用的单步法是不低于 p 阶的，则 k 步法(7.33)的整体截断误差为

$$\varepsilon_m=y(t_m)-y_m=O(h^p),\quad m\geqslant k$$

此定理的证明参见文献[5]第 36～37 页。

常用的线性 k 步法都满足相容性条件和根条件。

7.3.4　绝对稳定性

线性 k 步法(7.33)用于求解模型方程 $y'=\lambda y$，得到计算公式

$$\sum_{j=0}^{k}(\alpha_j-\mu\beta_j)y_{n+j}=0\quad(n=0,1,\cdots,M-k)\tag{7.48}$$

其中 $\mu=h\lambda$。设开始值 $y_0,y_1,\cdots,y_{k-1}$ 分别有扰动 $e_0,e_1,\cdots,e_{k-1}$，实际数值为 $\tilde{y}_0,\tilde{y}_1,\cdots,\tilde{y}_{k-1}$，其中 $e_m=y_m-\tilde{y}_m(m=0,1,\cdots,k-1)$。从这些带有扰动的开始值出发，由式(7.48)计算 $y_m(m\geqslant k)$，只能得到 y_m 的近似值 $\tilde{y}_m(m\geqslant k)$。假定用式(7.48)进行计算时没有任何舍入误差，误差 $e_m=y_m-\tilde{y}_m(m\geqslant k)$ 完全是由开始值的误差引起的，于是有

$$\sum_{j=0}^{k}(\alpha_j-\mu\beta_j)\tilde{y}_{n+j}=0\quad(n=0,1,\cdots,M-k)\tag{7.49}$$

将式(7.48)与式(7.49)相减，得到误差方程

$$\sum_{j=0}^{k}(\alpha_j-\mu\beta_j)e_{n+j}=0\quad(n=0,1,\cdots,M-k)\tag{7.50}$$

这是一个常系数线性齐次差分方程，它的特征方程为

$$\sum_{j=0}^{k}(\alpha_j-\mu\beta_j)\xi^j=0\tag{7.51}$$

或写成

$$\rho(\xi)-\mu\sigma(\xi)=0\tag{7.52}$$

代数方程(7.51)或方程(7.52)也称为线性 k 步法(7.33)的特征方程。设 ξ_i 是特征方程(7.51)的 r_i 重根($i=1,2,\cdots,s;r_1+r_2+\cdots+r_s=k$)，其中 $\xi_1,\xi_2,\cdots,\xi_s$ 互异，则差分方程(7.50)的通解为

$$e_m=\sum_{i=1}^{s}\sum_{l=1}^{r_i}c_{il}m^{l-1}\xi_i^m$$

(证明从略)。由初始值 $e_0,e_1,\cdots,e_{k-1}$ 可定出 k 个常数 c_{il}，并且每个 c_{il} 都是 $e_0,e_1,\cdots,e_{k-1}$ 的线性组合。由不等式

$$|e_m|\leqslant\sum_{i=1}^{s}\sum_{l=1}^{r_i}|c_{il}|m^{l-1}|\xi_i|^m$$

可知，当特征方程(7.51)的所有根 ξ_i 都满足 $|\xi_i|<1$ 时，就有 $\lim\limits_{m\to\infty}e_m=0$。

定义 对指定的 $\mu=h\lambda$，如果特征方程(7.51)的所有根 ξ_i 按模小于 1，则称线性 k 步法(7.33)关于此 μ 是绝对稳定的。若在 μ 复平面上有一区域 G，对一切 $\mu\in G$，k 步法(7.33)绝对稳定，则 G 称为 k 步法(7.33)的绝对稳定区域。

据定义，k 步法(7.33)的绝对稳定区域为

$$G=\left\{\mu\,\middle|\,|\xi|<1,\sum_{j=0}^{k}(\alpha_j-\mu\beta_j)\xi^j=0\right\}$$

与单步法的情形相类似，如果求解的是一般非线性微分方程(7.1)，则应视 $\lambda=\frac{\partial f}{\partial y}$。这时 λ 可能是变化的，但只要在求解区间 $t_0\leqslant t\leqslant T$ 内，$\mu=h\lambda$ 始终属于绝对稳定区域 G，则对此微分方程而言，k 步法(7.33)是绝对稳定的。

例 6 求向后 Euler 法

$$y_{n+1}=y_n+hf_{n+1}$$

的绝对稳定区域。

解 特征方程

$$\rho(\xi)-\mu\sigma(\xi)=\xi-1-\mu\xi=0$$

的根为 $\xi=\frac{1}{1-\mu}$，故向后 Euler 法的绝对稳定区域是 $|1-\mu|>1$，见图 7-2。

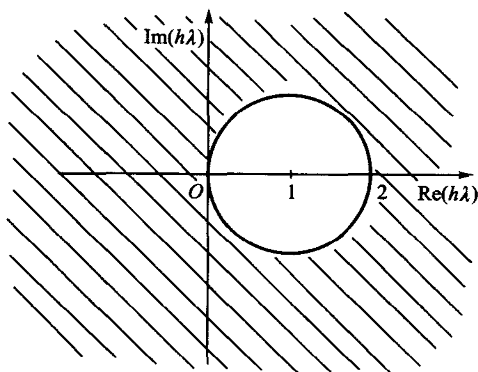


图 7-2 向后 Euler 法的绝对稳定区域

例 7 求梯形法

$$y_{n+1}=y_n+\frac{h}{2}(f_n+f_{n+1})$$

的绝对稳定区域。

解 特征方程

$$\rho(\xi)-\mu\sigma(\xi)=\xi-1-\frac{\mu}{2}(\xi+1)=0$$

的根为 $\xi=\frac{2+\mu}{2-\mu}$，因而梯形法的绝对稳定区域是 $\left|\frac{2+\mu}{2-\mu}\right|<1$，即 $\mathrm{Re}\mu<0$（μ 复平面上的整个左半平面，见图 7-3）。

例 8 求显式 Milne 方法

$$y_{n+4}=y_n+\frac{4}{3}h(2f_{n+1}-f_{n+2}+2f_{n+3})$$

的绝对稳定区域。

解　特征方程为

$$\xi^4 - \frac{8}{3}\mu\xi^3 + \frac{4}{3}\mu\xi^2 - \frac{8}{3}\mu\xi - 1 = 0$$

由于系数 $\alpha_4 = 1, \alpha_0 = -1$，所以此方程的所有根 $\xi_1, \xi_2, \xi_3, \xi_4$ 满足

$$|\xi_1\ \xi_2\ \xi_3\ \xi_4| = \left|\frac{\alpha_0}{\alpha_4}\right| = 1$$

因而必存在某个根 ξ_i，$|\xi_i| \geqslant 1$。可见，显式 Milne 方法是一个恒不绝对稳定的方法。

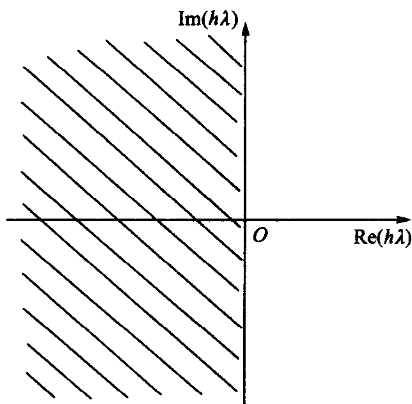


图 7-3　梯形法的绝对稳定区域

用直接求出特征方程的根去确定线性多步法的绝对稳定区域，在大多数情况下是行不通的。一般情况下可使用边界轨迹法。

根据绝对稳定性的定义，只要 μ 位于 μ 复平面的绝对稳定区域 G 内，特征方程(7.52)的根就位于 ξ 复平面的单位圆域 $|\xi| < 1$ 之内。如果特征方程(7.52)的根位于单位圆周 $|\xi| = 1$（即 $\xi = e^{i\theta}$）上，则满足方程

$$\rho(e^{i\theta}) - \mu\sigma(e^{i\theta}) = 0, \quad 0 \leqslant \theta \leqslant 2\pi$$

的 μ 就位于绝对稳定区域 G 的边界 Γ 上。因此，当 $\sigma(e^{i\theta}) \neq 0 (0 \leqslant \theta \leqslant 2\pi)$ 时，边界 Γ 的轨迹由

$$\mu(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \quad 0 \leqslant \theta \leqslant 2\pi$$

即参数方程

$$\begin{cases} x(\theta) = \mathrm{Re}\,\mu(\theta) \\ y(\theta) = \mathrm{Im}\,\mu(\theta) \\ \qquad 0 \leqslant \theta \leqslant 2\pi \end{cases}$$

给出。当 θ 从 0 变化到 2π 时，$\mu(\theta)$ 在 μ 复平面上就描出边界曲线 Γ。该曲线把 μ 复平面分为两个区域，如果在其中一个区域的内部存在一点 μ_1，使方程

$$\rho(\xi) - \mu_1\sigma(\xi) = 0$$

有按模大于 1 的根，则该区域就不是绝对稳定区域，而另一个区域是绝对稳定区域。这就是边界轨迹法。

以下的定理常用来配合边界轨迹法。

定理 7.5　实系数 k 次代数方程 $\sum_{j=0}^{k}\alpha_j\xi^j = 0 (\alpha_k > 0)$ 的所有根按模小于 1 的必要条件是下列不等式同时成立：

$$|\alpha_0| < \alpha_k, \quad \sum_{j=0}^{k}\alpha_j > 0, \quad \sum_{j=0}^{k}(-1)^{k-j}\alpha_j > 0$$

而对于 $k=2$ 的情形，这些不等式就成为充分必要条件。

此定理的证明参见文献[12]。

例 9　求三步隐式 Adams 方法

$$y_{n+3} = y_{n+2} + \frac{h}{24}(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n)$$

的绝对稳定区域。

解 $\rho(\xi)=\xi^3-\xi^2,\quad \sigma(\xi)=\frac{1}{24}(9\xi^3+19\xi^2-5\xi+1)$

$$\mu(\theta)=\frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}=\frac{24(e^{3i\theta}-e^{2i\theta})}{9e^{3i\theta}+19e^{2i\theta}-5e^{i\theta}+1}=x(\theta)+iy(\theta)$$

绝对稳定区域 G 的边界 Γ 就是下列参数方程所表示的曲线：

$$\begin{cases} x(\theta)=24(-10+15\cos\theta-6\cos 2\theta+\cos 3\theta)/r(\theta) \\ y(\theta)=24(33\sin\theta-6\sin 2\theta+\sin 3\theta)/r(\theta) \end{cases}$$

其中 $r(\theta)=468+142\cos\theta-52\cos 2\theta+18\cos 3\theta$。

显然，$x(-\theta)=x(\theta)$，$y(-\theta)=-y(\theta)$，所以边界 Γ 关于实轴对称。列表 7-6 计算 x,y 的值。

表 7-6 例 9 边界曲线上点的坐标

θ	0°	30°	60°	90°	120°	150°	180°
x	0	−0.000	−0.002	−0.185	−0.735	−1.955	−3.000
y	0	0.523	1.026	1.477	1.838	1.707	0

在 μ 平面上描点、连线，利用对称性就得到边界曲线，见图 7-4。

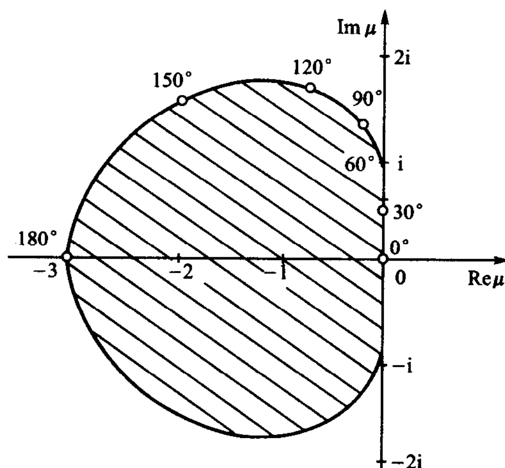


图 7-4 三步隐式 Adams 方法的绝对稳定区域

为判明绝对稳定区域是在闭曲线 Γ 所围之域内还是域外，取$\mu=1$，则相应的特征方程是

$$\rho(\xi)-\sigma(\xi)=15\xi^3-43\xi^2+5\xi-1=0$$

由于它的系数之和 $15-43+5-1<0$，根据定理 7.5 以及 $\mu=1$ 不在 Γ 上，可知该特征方程必有模大于 1 的根。所以，闭曲线 Γ 所围之域内是绝对稳定区域。

利用边界轨迹法可求出 k 从 1 到 6 的 Gear 方法的绝对稳定区域，见图 7-5 和图 7-6。区域的边界都是封闭曲线，绝对稳定区域是在相应闭曲线所围之域的外部。

当 λ 为实数时，$\mu=h\lambda$ 也是实数，实数 μ 在绝对稳定区域内的取值范围就是绝对稳定区间。使用边界轨迹法可直接求出各种线性多步法的绝对稳定区间。下面就是一些线性多步法的绝对稳定区间。

三步显式 Adams 方法(7.37)： $-\frac{6}{11}<\mu<0$

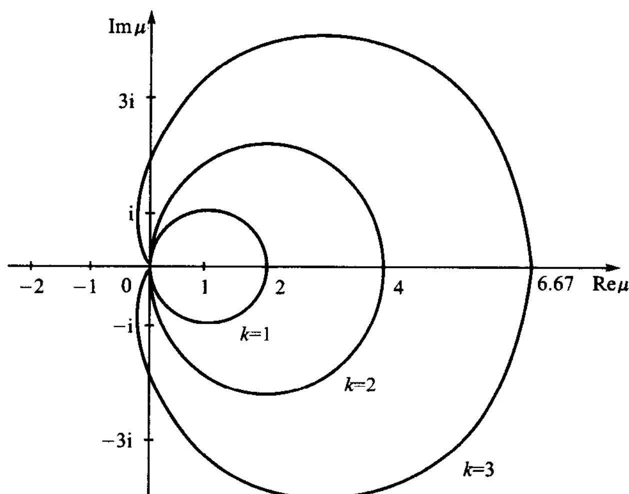


图 7-5　Gear 方法绝对稳定区域边界线($k=1,2,3$)

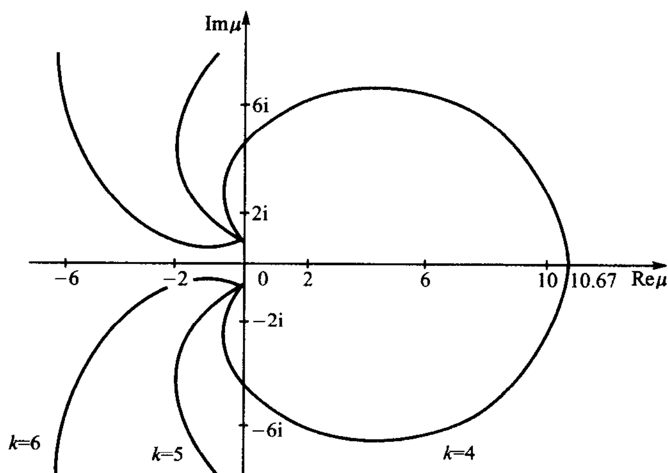


图 7-6　Gear 方法绝对稳定区域边界线($k=4,5,6$)

三步隐式 Adams 方法(7.38)：　$-3<\mu<0$

四步显式 Adams 方法(7.39)：　$-0.3<\mu<0$

Hamming 公式(7.41)：　$-\dfrac{8}{3}<\mu<0$

梯形法(7.42)：　$-\infty<\mu<0$

利用定理 7.5 中 $k=2$ 的情形可直接求出线性二步法的绝对稳定区间，因为此时 $\mu=h\lambda$ 是实数，线性二步法的特征方程是实系数二次代数方程。

例 10　求二步隐式 Adams 方法式(7.36)，即

$$y_{n+2}=y_{n+1}+\frac{h}{12}(5f_{n+2}+8f_{n+1}-f_n)$$

的绝对稳定区间。

解 特征方程为

$$\left(1-\frac{5}{12}\mu\right)\xi^2-\left(1+\frac{8}{12}\mu\right)\xi+\frac{\mu}{12}=0$$

设

$$1-\frac{5}{12}\mu>0 \quad 即 \quad \mu<\frac{12}{5}$$

由定理 7.5 知,特征方程的两个根之模皆小于 1 的充分必要条件是

$$\begin{cases}\left|\dfrac{\mu}{12}\right|<1-\dfrac{5}{12}\mu & 即 \quad \mu<2\\ 1-\dfrac{5}{12}\mu-\left(1+\dfrac{8}{12}\mu\right)+\dfrac{\mu}{12}>0 & 即 \quad \mu<0\\ 1-\dfrac{5}{12}\mu+\left(1+\dfrac{8}{12}\mu\right)+\dfrac{\mu}{12}>0 & 即 \quad \mu>-6\end{cases}$$

上述 μ 的四个取值范围的交集是 $-6<\mu<0$。

设

$$1-\frac{5}{12}\mu<0 \quad 即 \quad \mu>\frac{12}{5}$$

特征方程写成

$$\left(\frac{5}{12}\mu-1\right)\xi^2+\left(1+\frac{8}{12}\mu\right)\xi-\frac{\mu}{12}=0$$

再由定理 7.5 中的三个条件分别得

$$\mu>3,\quad \mu>0,\quad \mu<-6$$

上述 μ 的四个取值范围无交集。

设

$$1-\frac{5}{12}\mu=0 \quad 即 \quad \mu=\frac{12}{5}$$

这时特征方程只有一个根 ξ:$\xi=\frac{1}{13}$,$|\xi|<1$。

综上所述,二步隐式 Adams 方法(7.36)的绝对稳定区间为

$$-6<\mu<0 \quad 和 \quad \mu=\frac{12}{5}$$

定义 一个求解初值问题(7.1)、(7.2)的数值解法称为是 $A(\alpha)$-稳定的,如果它的绝对稳定区域包含了无限楔形区域

$$W_\alpha=\{\mu\,|-\alpha<\pi-\arg\mu<\alpha\}$$

其中 $0<\alpha\leqslant\frac{\pi}{2}$ 见图 7-7。

从图 7-5 和图 7-6 可看出,k 从 1 到 6 的 Gear 方法都是 $A(\alpha)$-稳定的,其中每个 α 的最大值 α_{max} 如下:

k	1	2	3	4	5	6
α_{max}	90°	90°	86°54′	73°14′	51°50′	18°47′

对于微分方程(7.1),只要 $\lambda=\frac{\partial f}{\partial y}$ 位于 μ 复平面的区域 W_α 内,那么单从数值稳定性考虑,

使用 $A(\alpha)$-稳定的数值解法求解时，步长 h 可不受限制，因为无论 $h>0$ 为何值，总有 $\mu=h\lambda\in W_\alpha$。

定义　一个求解初值问题(7.1)、(7.2)的数值解法称为是 A-稳定的，如果它的绝对稳定区域包含了 μ 复平面的整个左半平面。

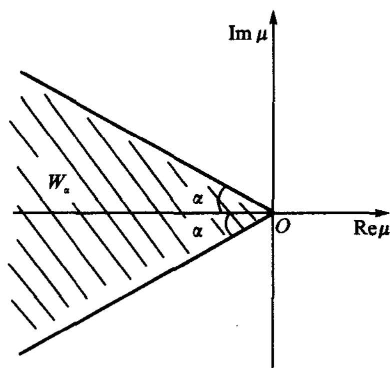


图 7-7　$A(\alpha)$-稳定的楔形区域

$A\left(\frac{\pi}{2}\right)$-稳定就是 A-稳定。例如，向后 Euler 法、梯形法和二步 Gear 方法都是 A-稳定的方法。对于固有稳定的微分方程，即 $\mathrm{Re}\,\frac{\partial f}{\partial y}<0$ 的微分方程(7.1)，当使用 A-稳定的数值解法求解时，单从数值稳定性考虑，步长 h 不受任何限制。

经证明，有这样的结论：显式线性多步法不可能是 A-稳定的；A-稳定的线性多步法的阶不超过 2，而且在二阶 A-稳定的线性多步法中梯形法的误差常数最小。

7.4　步长的选择

使用数值解法求解初值问题(7.1)、(7.2)，选择步长 h 是十分重要的问题。原则上，步长 h 应由两个条件决定，一是要使求解过程绝对稳定，二是使每个节点处的整体截断误差 ε_{n+1} 按模不超过给定的界限。

对于第一个条件，只要根据所给初值问题和所用方法的绝对稳定区域就可确定(或大致确定)h 所受的限制。根据这个限制和实际需要的离散化程度，先确定节点步长 $h_0>0$，从而节点

$$t_n = t_0 + nh_0 \quad (n = 0,1,\cdots,M) \tag{7.53}$$

被确定。如果用此不变的步长 h_0 求出数值解 $y_1,y_2,\cdots,y_M$，那么，所得的结果是否满足精度要求是很难检验的，这是因为整体截断误差 ε_{n+1} 的表达式中的首项系数 $c(t_{n+1})$ 很难估计。由于同样的理由，也不能预先根据 ε_{n+1} 的表达式确定步长 h_0 的范围。

在实际计算中，常采用外推法估计误差以及自动选择步长的计算方案计算给定节点(7.53)上的数值解 $y_n(n=1,2,\cdots,M)$。

设所用的是 p 阶单步法，从 t_n 开始采用步长 h 计算出 t_{n+1} 处的数值解为 $y_{n+1,h}$，又从 t_n 开始采用步长 $\frac{h}{2}$ 计算出 t_{n+1} 处的数值解为 $y_{n+1,h/2}$。$y_{n+1,h}$ 和 $y_{n+1,h/2}$ 的整体截断误差分别为

$$y(t_{n+1}) - y_{n+1,h} = c(t_{n+1})h^p + O(h^{p+1}) \tag{7.54}$$

$$y(t_{n+1}) - y_{n+1,h/2} = c(t_{n+1})\left(\frac{h}{2}\right)^p + O(h^{p+1}) \tag{7.55}$$

两式相减，得

$$y_{n+1,h} - y_{n+1,h/2} = \left(\frac{1}{2^p} - 1\right)c(t_{n+1})h^p + O(h^{p+1}) \tag{7.56}$$

由式(7.54)和式(7.56)消去 $c(t_{n+1})h^p$，得

$$y(t_{n+1}) - y_{n+1,h} = \frac{2^p}{1-2^p}(y_{n+1,h} - y_{n+1,h/2}) + O(h^{p+1})$$

由此看出，$y_{n+1,h}$的整体截断误差可近似计算为

$$\varepsilon_{n+1,h} = y(t_{n+1}) - y_{n+1,h} \approx \frac{2^p}{1-2^p}(y_{n+1,h} - y_{n+1,h/2}) \tag{7.57}$$

h 越小近似程度越高。式(7.57)称为用外推法估计误差，也称为事后误差估计式。

在通用的微分方程数值解法程序中，通常包括外推法估计误差及自动选择步长的过程。关于自动选择步长，其具体做法如下：设已按选定的某个步长 $h(h=h_0/2^m$，整数 $m\geqslant 0)$算出满足精度要求的数值解 y_n，然后以步长 h 及$\frac{h}{2}$分别计算 $y_{n+1,h}$和 $y_{n+1,h/2}$，并用式(7.57)估计 $y_{n+1,h}$的误差。如果

$$|\varepsilon_{n+1,h}| < \delta \quad (\delta \text{ 为给定精度}) \tag{7.58}$$

且相差不大，则取 $y_{n+1}=y_{n+1,h}$，接着用同样的步长 h 计算节点 t_{n+2}处的值 y_{n+2}；如果 $\varepsilon_{n+1,h}$不满足式(7.58)，则应取$\frac{h}{2}$为新步长重新计算，直到外推法估计的误差小于 δ 为止，并在下一步的计算中把这样选择的步长作为新步长；如果$|\varepsilon_{n+1,h}|$远远小于 δ，这表明步长过小，应将步长加一倍，以 $2h\leqslant h_0$ 作为新步长计算 y_{n+2}。如此继续下去。一开始计算 y_1 时，可取 $h=h_0$。

7.5 常微分方程组与刚性问题

7.5.1 常微分方程组初值问题的数值解法

设有一阶常微分方程组初值问题

$$\begin{cases} y'_i = f_i(t, y_1, y_2, \cdots, y_s) \quad (i=1,2,\cdots,s) \quad t_0\leqslant t\leqslant T & (7.59) \\ y_i(t_0) = y_{i0} \quad (i=1,2,\cdots,s) & (7.60) \end{cases}$$

未知函数 $y_1(t),\cdots,y_s(t)$的个数 s 称为微分方程组(7.59)的维数。记

$$\boldsymbol{y} = (y_1, y_2, \cdots, y_s)^{\mathrm{T}}, \quad \boldsymbol{y}_0 = (y_{10}, y_{20}, \cdots, y_{s0})^{\mathrm{T}}$$

$$\boldsymbol{f} = (f_1, f_2, \cdots, f_s)^{\mathrm{T}}$$

则初值问题(7.59)、(7.60)可表示为向量形式

$$\begin{cases} \boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad t_0 \leqslant t \leqslant T & (7.61) \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 & (7.62) \end{cases}$$

只须把一维情形的所有数值解法推广到 s 维情形就可用于求解 s 维初值问题(7.61)、(7.62)。在一维情形下得出的所有数值解法都可看做维数未定，它们的维数只由求解对象确定。例如，用 Euler 法(7.13)求解下列二维初值问题

$$\begin{cases} u' = f(t,u,v), \quad u(t_0) = u_0 \\ v' = g(t,u,v), \quad v(t_0) = v_0 \end{cases} \quad t_0 \leqslant t \leqslant T$$

则 Euler 法(7.13)就是二维的，求解公式为

$$\begin{bmatrix} u_{n+1} \\ v_{n+1} \end{bmatrix} = \begin{bmatrix} u_n \\ v_n \end{bmatrix} + h\begin{bmatrix} f(t_n, u_n, v_n) \\ g(t_n, u_n, v_n) \end{bmatrix} \quad (n = 0,1,\cdots,M-1)$$

若用改进的 Euler 法(7.19)求解，则求解公式为

$$\begin{cases} \boldsymbol{k}_1 = \begin{bmatrix} k_{11} \\ k_{21} \end{bmatrix} = \begin{bmatrix} f(t_n, u_n, v_n) \\ g(t_n, u_n, v_n) \end{bmatrix} \\ \boldsymbol{k}_2 = \begin{bmatrix} k_{12} \\ k_{22} \end{bmatrix} = \begin{bmatrix} f(t_n + h, u_n + hk_{11}, v_n + hk_{21}) \\ g(t_n + h, u_n + hk_{11}, v_n + hk_{21}) \end{bmatrix} \\ \begin{bmatrix} u_{n+1} \\ v_{n+1} \end{bmatrix} = \begin{bmatrix} u_n \\ v_n \end{bmatrix} + \dfrac{h}{2} \begin{bmatrix} k_{11} + k_{12} \\ k_{21} + k_{22} \end{bmatrix} \quad (n = 0, 1, \cdots, M-1) \end{cases}$$

对于一维 m 阶常微分方程初值问题

$$\begin{cases} y^{(m)} = f(t, y, y', \cdots, y^{(m-1)}), \quad t_0 \leqslant t \leqslant T \\ y^{(i)}(t_0) = y_0^{(i)} \quad (i = 0, 1, \cdots, m-1) \end{cases}$$

可通过变量代换化为 m 维一阶常微分方程组初值问题，再用数值解法求解。

例如，对于二阶常微分方程初值问题

$$\begin{cases} y'' = f(t, y, y'), \quad t_0 \leqslant t \leqslant T \\ y(t_0) = y_0, \quad y'(t_0) = y'_0 \end{cases}$$

可令 $z = y'(t)$，则把此初值问题化为下列的一阶常微分方程组

$$\begin{cases} y' = z, & y(t_0) = y_0 \\ z' = f(t, y, z), & z(t_0) = y'_0 \end{cases}$$

若用四阶 R－K 法(7.27)求解，则计算公式为

$$\begin{cases} \boldsymbol{k}_1 = \begin{bmatrix} k_{11} \\ k_{21} \end{bmatrix} = \begin{bmatrix} z_n \\ f(t_n, y_n, z_n) \end{bmatrix} \\ \boldsymbol{k}_2 = \begin{bmatrix} k_{12} \\ k_{22} \end{bmatrix} = \begin{bmatrix} z_n + \dfrac{h}{2} k_{21} \\ f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2} k_{11}, z_n + \dfrac{h}{2} k_{21}\right) \end{bmatrix} \\ \boldsymbol{k}_3 = \begin{bmatrix} k_{13} \\ k_{23} \end{bmatrix} = \begin{bmatrix} z_n + \dfrac{h}{2} k_{22} \\ f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2} k_{12}, z_n + \dfrac{h}{2} k_{22}\right) \end{bmatrix} \\ \boldsymbol{k}_4 = \begin{bmatrix} k_{14} \\ k_{24} \end{bmatrix} = \begin{bmatrix} z_n + hk_{23} \\ f(t_n + h, y_n + hk_{13}, z_n + hk_{23}) \end{bmatrix} \\ \begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} y_n \\ z_n \end{bmatrix} + \dfrac{h}{6} \begin{bmatrix} k_{11} + 2k_{12} + 2k_{13} + k_{14} \\ k_{21} + 2k_{22} + 2k_{23} + k_{24} \end{bmatrix} \\ \qquad (n = 0, 1, \cdots, M-1) \end{cases}$$

其中 $(y_0, z_0)^{\mathrm{T}} = (y_0, y'_0)^{\mathrm{T}}$。

例 11　试写出用中点公式(7.20)求解下列初值问题

$$\begin{cases} y''' = ty + y'y'', \quad 0 \leqslant t \leqslant 20 \\ y(0) = 1, \quad y'(0) = 0, \quad y''(0) = -2 \end{cases}$$

的计算公式。

解　把所给初值问题化为一阶微分方程组初值问题：

$$\begin{cases} x' = z, & x(0) = 0 \\ y' = x, & y(0) = 1 \qquad 0 \leqslant t \leqslant 20 \\ z' = ty + xz, & z(0) = -2 \end{cases}$$

求解的计算公式为

$$
\begin{cases}
(x_0, y_0, z_0)^{\mathrm{T}} = (0, 1, -2)^{\mathrm{T}} \\
\begin{bmatrix} k_{11} \\ k_{21} \\ k_{31} \end{bmatrix} = \begin{bmatrix} z_n \\ x_n \\ nhy_n + x_n z_n \end{bmatrix} \\
\begin{bmatrix} k_{12} \\ k_{22} \\ k_{32} \end{bmatrix} = \begin{bmatrix} z_n + 0.5hk_{31} \\ x_n + 0.5hk_{11} \\ (n+0.5)h(y_n + 0.5hk_{21}) + (x_n + 0.5hk_{11})(z_n + 0.5hk_{31}) \end{bmatrix} \\
\begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} + h \begin{bmatrix} k_{12} \\ k_{22} \\ k_{32} \end{bmatrix} \\
\qquad \left(n = 0, 1, \cdots, \left[\frac{20}{h}\right] - 1\right)
\end{cases}
$$

由一维情形建立的有关相容性和收敛性的概念和定理也适用于多维情形，此时须要把函数的绝对值换成函数向量的范数，例如函数向量 $\boldsymbol{\Phi}(t, \boldsymbol{y}, h)$ 关于向量 $\boldsymbol{y}$ 的 Lipschitz 条件的形式为

$$\|\boldsymbol{\Phi}(t, \boldsymbol{u}_1, h) - \boldsymbol{\Phi}(t, \boldsymbol{u}_2, h)\| \leqslant K \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|$$

在方程组的情形，定义绝对稳定性的模型方程是 s 维的线性微分方程组

$$\boldsymbol{y}' = \boldsymbol{\Lambda}\boldsymbol{y} \tag{7.63}$$

其中 $\boldsymbol{\Lambda}$ 是对角矩阵，表示为

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_s)$$

其中 $\lambda_i (i=1,2,\cdots,s)$ 皆为复常数。

定义 设步长为 $h>0$ 的单步法(7.6)用于求解模型方程(7.63)的初值问题，并设初始向量 $\boldsymbol{y}_0$ 有误差

$$\boldsymbol{e}_0 = (e_{10}, e_{20}, \cdots, e_{s0})^{\mathrm{T}}$$

如果在计算后面的 $\boldsymbol{y}_n$ 时，由 $\boldsymbol{e}_0$ 所引起的误差

$$\boldsymbol{e}_n = (e_{1n}, e_{2n}, \cdots, e_{sn})^{\mathrm{T}}$$

满足

$$\lim_{n \to \infty} \boldsymbol{e}_n = \boldsymbol{0}$$

则称单步法(7.6)对于所用的步长 h 和 $\lambda_i (i=1,2,\cdots,s)$ 是绝对稳定的。

s 维线性多步法的绝对稳定性定义与一维情形完全相同。

一维情形的各种数值解法的绝对稳定区域也是 s 维情形下的绝对稳定区域。

例如，用 s 维 Euler 法求解模型方程(7.63)的初值问题。得计算公式

$$\boldsymbol{y}_{n+1} = (\boldsymbol{I} + h\boldsymbol{\Lambda})\boldsymbol{y}_n$$

其中 $\boldsymbol{I}$ 是 $s \times s$ 单位矩阵。设实际参加运算的是 $\tilde{\boldsymbol{y}}_n$，误差为 $\boldsymbol{e}_n = \boldsymbol{y}_n - \tilde{\boldsymbol{y}}_n$，于是得

$$\tilde{\boldsymbol{y}}_{n+1} = (\boldsymbol{I} + h\boldsymbol{\Lambda})\tilde{\boldsymbol{y}}_n$$

上面两式相减得

$$\boldsymbol{e}_{n+1} = (\boldsymbol{I} + h\boldsymbol{\Lambda})\boldsymbol{e}_n$$

因而有

$$e_{n+1} = (I + h\Lambda)^{n+1} e_0$$

$$e_{i,n+1} = (1 + h\lambda_i)^{n+1} e_{i0} \quad (i = 1,2,\cdots,s)$$

当且仅当

$$|1 + h\lambda_i| < 1 \quad (i = 1,2,\cdots,s)$$

时，有

$$\lim_{n\to\infty} e_n = 0$$

由此可知，s 维 Euler 法的绝对稳定区域仍是

$$|1 + h\lambda| < 1 \tag{7.64}$$

但是，在使用时，λ 要取遍 $\lambda_1,\lambda_2,\cdots,\lambda_s$。

设方程组(7.61)是一般常系数线性非齐次微分方程组

$$y' = Ay + \psi(t) \tag{7.65}$$

其中 A 是 $s\times s$ 常数矩阵，并且有 s 个线性无关的特征向量，又

$$\psi(t) = (\psi_1(t),\psi_2(t),\cdots,\psi_s(t))^{\mathrm{T}}$$

用 Euler 法求解式(7.65)的初值问题，其计算公式是

$$y_{n+1} = y_n + h[Ay_n + \psi(t_n)] = (I + hA)y_n + h\psi(t_n)$$

设 $e_n = y_n - \tilde{y}_n$，因只考虑 e_n 的传播，故有

$$e_{n+1} = (I + hA)e_n \tag{7.66}$$

因 A 有 s 个线性无关的特征向量，故存在非奇异矩阵 P，使

$$P^{-1}AP = \Lambda = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_s)$$

或

$$A = P\Lambda P^{-1}$$

其中 $\lambda_i(i=1,2,\cdots,s)$是 A 的特征值。于是，误差方程(7.66)成为

$$e_{n+1} = (I + hP\Lambda P^{-1})e_n = P(I + h\Lambda)P^{-1}e_n$$

或

$$\eta_{n+1} = (I + h\Lambda)\eta_n$$

其中 $\eta_n = P^{-1}e_n$，并且当 $\lim\limits_{n\to\infty}\eta_n = 0$ 时也有 $\lim\limits_{n\to\infty} e_n = 0$。由此可知，当用 Euler 法求解方程组(7.65)的初值问题须考察绝对稳定性时，不等式(7.64)中的 λ 要取遍矩阵 A 的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_s$。同理可知，若用其他方法求解式(7.65)的初值问题须考察其绝对稳定性时，那么，表示绝对稳定区域的不等式中的 λ，也要取遍矩阵 A 的所有特征值。

设初值问题(7.61)、(7.62)中的方程组是一般微分方程组，并设 f 关于 y 的 Jacobi 矩阵(简称微分方程组的 Jacobi 矩阵)

$$\frac{\partial f}{\partial y} = \left[\frac{\partial f_i}{\partial y_j}\right]_{s\times s}$$

在区间 $t_0\leqslant t\leqslant T$ 内有 s 个线性无关的特征向量，那么，在考察一个方法求解式(7.61)的绝对稳定性时，该方法的绝对稳定区域中的 λ 要取遍矩阵$\dfrac{\partial f}{\partial y}$的所有特征值。此时各个特征值 λ 可能是变量，选取 h 时，要使 $h\lambda$ 始终位于所用方法的绝对稳定区域内。

例 12　用四级四阶 R - K 方法(7.27)求解初值问题

$$\begin{cases} x' = -1.5x + 0.5y - 0.5z, & x(0) = x_0 \\ y' = 1.5x - 2.5y - 1.5z, & y(0) = y_0 \qquad 0 \leqslant t \leqslant 100 \\ z' = x - y - 3z, & z(0) = z_0 \end{cases}$$

时，由绝对稳定性对步长 h 有何限制？

解

$$\boldsymbol{A} = \begin{bmatrix} -1.5 & 0.5 & -0.5 \\ 1.5 & -2.5 & -1.5 \\ 1 & -1 & -3 \end{bmatrix}$$

因

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = (\lambda + 2)(\lambda + 1)(\lambda + 4)$$

故 $\boldsymbol{A}$ 的特征值为 $\lambda_1 = -2, \lambda_2 = -1, \lambda_3 = -4$。四级四阶 R－K 方法(7.27)的绝对稳定区间为

$$-2.78 < h\lambda < 0$$

由 λ_1, λ_2 和 λ_3 依次代入上式中的 λ，分别得

$$0 < h < 1.39, \quad 0 < h < 2.78, \quad 0 < h < 0.695$$

因此，对 h 的限制应是上述三个不等式的公共部分 $0<h<0.695$。

例 13 用 Euler 法求解初值问题

$$\begin{cases} x' = ty - x + \sqrt{t}, & x(0) = x_0 \\ y' = -2tx - 5y + \sin t, & y(0) = y_0 \end{cases} \quad 0 \leqslant t \leqslant 2$$

时，由绝对稳定性对步长有何限制？

解

$$\boldsymbol{A} = \begin{bmatrix} -1 & t \\ -2t & -5 \end{bmatrix}$$

由 $\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0$ 解得 $\boldsymbol{A}$ 的特征值为

$$\lambda_{1,2} = -3 \pm \sqrt{4 - 2t^2}$$

当 $0 \leqslant t \leqslant \sqrt{2}$ 时，λ_1, λ_2 都是实数。对于 λ_1，由绝对稳定区间 $-2<h\lambda<0$，得 h 的取值范围

$$0 < h < \min_{0 \leqslant t \leqslant \sqrt{2}} \frac{2}{3 - \sqrt{4 - 2t^2}} = \frac{2}{3}$$

对于 λ_2，又得

$$0 < h < \min_{0 \leqslant t \leqslant \sqrt{2}} \frac{2}{3 + \sqrt{4 - 2t^2}} = \frac{2}{5}$$

易知，当 $0 \leqslant t \leqslant \sqrt{2}$，对 h 的限制为 $0<h<\frac{2}{5}$。

当 $\sqrt{2}<t\leqslant 2$ 时，$\lambda_{1,2} = -3 \pm \mathrm{i}\sqrt{2t^2 - 4}$，由

$$|1 + h\lambda_{1,2}| < 1$$

解得 h 的取值范围

$$0 < h < \min_{\sqrt{2} < t \leqslant 2} \frac{6}{2t^2 + 5} = \frac{6}{13}$$

如果在整个求解区间 $0 \leqslant t \leqslant 2$ 内用不变的步长，则应使 h 满足

$$0<h<\frac{2}{5}$$

7.5.2 刚性问题

设有一阶微分方程组初值问题

$$\begin{cases} u' = 998u + 1\,998v, & u(0) = 1 \\ v' = -999u - 1\,999v, & v(0) = 0 \end{cases} \qquad 0 \leqslant t \leqslant T$$

矩阵

$$\boldsymbol{A} = \begin{bmatrix} 998 & 1\,998 \\ -999 & -1\,999 \end{bmatrix}$$

的特征值为 $\lambda_1=-1,\lambda_2=-1\,000$。此初值问题描述了一个线性系统在初始干扰 $u(0)=1$，$v(0)=0$的影响下，状态$(u,v)^{\mathrm{T}}$ 的变化情况，其解为

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2\mathrm{e}^{-t} \\ -\mathrm{e}^{-t} \end{bmatrix} + \begin{bmatrix} -\mathrm{e}^{-1\,000t} \\ \mathrm{e}^{-1\,000t} \end{bmatrix}$$

随着时间的增加，$(u,v)^{\mathrm{T}}$ 将恢复原来的稳定状态$(0,0)^{\mathrm{T}}$。从解的结构可看到，状态$(u,v)^{\mathrm{T}}$ 由两部分叠加而成。相应于 $\lambda_1=-1$ 的部分为$(2\mathrm{e}^{-t},-\mathrm{e}^{-t})^{\mathrm{T}}$，相应于 $\lambda_2=-1\,000$ 的部分是$(-\mathrm{e}^{-1\,000t},\mathrm{e}^{-1\,000t})^{\mathrm{T}}$。当 t 增加时，相对于第一部分，第二部分衰减得很快，称它为快变部分，而第一部分称为慢变部分。比值

$$r = \max_i |\mathrm{Re}\lambda_i| / \min_i |\mathrm{Re}\lambda_i|$$

越大，快慢的差别也越大。

现在考虑用数值解法求解，将会看到，由于比值 r 大而给选择步长 h 造成困难。假定使用 Euler 法求解，为使求解过程绝对稳定，所用步长 h 应满足$|1+h\lambda_1|<1$ 和$|1+h\lambda_2|<1$。由前者得 $0<h<2$，由后者得 $0<h<0.002$，显然应使 $0<h<0.002$，也就是步长 h 由$\max\limits_i|\mathrm{Re}\lambda_i|$决定，并且必然是最小的。由于所用步长小，而$\min\limits_i|\mathrm{Re}\lambda_i|$也小，因而要想把慢变部分计算到稳态则所需的步数就多，并且会使计算时间长得无法接受。但是，如果步长 h 由 λ_1 决定，则要引起计算误差的积累，使计算结果完全丧失真实性。

上述现象是在由不同刚度(stiffness)的元件所控制的系统中出现的，所以具有这种现象的微分方程组叫做刚性方程组或 stiff 方程组，也叫病态方程组。研究刚性方程组的刚性程度和求解问题就叫刚性问题或 stiff 问题。

定义　s 维常系数线性微分方程组

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{\psi}(t)$$

称为刚性方程组，如果矩阵 $\boldsymbol{A}$ 的特征值 $\lambda_i(i=1,2,\cdots,s)$满足条件

(1) $\mathrm{Re}\lambda_i<0(i=1,2,\cdots,s)$；

(2) $\max\limits_i|\mathrm{Re}\lambda_i|\gg\min\limits_i|\mathrm{Re}\lambda_i|$。

并称比值

$$r = \max_i |\mathrm{Re}\lambda_i| / \min_i |\mathrm{Re}\lambda_i|$$

为刚性比。

对于一般 s 维微分方程组(7.61)，如果 Jacobi 矩阵$\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}$的特征值 $\lambda_i(i=1,2,\cdots,s)$在区间

$t_0 \leqslant t \leqslant T$上满足上述定义中的条件，那么，方程组(7.61)也称为刚性方程组。

由于刚性方程组是固有稳定的微分方程组，即 $\mathrm{Re}\lambda_i < 0 (i=1,2,\cdots,s)$，所以，$\lambda_i$ 总位于 $\mu = h\lambda$复平面的左半平面。如果使用 A-稳定的方法求解，则对任意的步长 $h>0$，$\mu = h\lambda_i (i=1,2,\cdots,s)$总位于方法的绝对稳定区域内。此时，只须从截断误差的控制考虑步长的选择。如果使用 $A(\alpha)$-稳定的方法求解，那么，只要所有的 λ_i 都位于所用方法的 W_α 区域内，就总有 $h\lambda_i \in W_\alpha$，而不论 $h>0$ 为何值。此时，同样只须从截断误差的控制考虑步长的选择。因此，A-稳定或 $A(\alpha)$-稳定的方法都适用于求解刚性方程组。梯形法以及各种 Gear 方法都是求解刚性方程组常用的方法。

到目前为止，A-稳定和 $A(\alpha)$-稳定的方法都是隐式方法。因此，对 s 维的刚性方程组，应用隐式线性 k 步法求解时，每计算一个 y_{n+k}都要用迭代法求解一个 s 元非线性方程组，通常采用 Newton 法或离散 Newton 法求解。

在实际应用中，还有一类隐式非线性单步法可用于求解刚性方程组，就是隐式 Runge - Kutta 方法，简称隐式 R - K 方法。

N 级隐式 R - K 方法的一般形式是

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{N} c_i k_i \\ k_i = f\left(t_n + \alpha_i h, y_n + \sum_{j=1}^{N} b_{ij} k_j\right) \quad (i=1,2,\cdots,N) \end{cases}$$

下面是常用的三个隐式 R - K 方法。

一级二阶公式：

$$\begin{cases} y_{n+1} = y_n + hk_1 \\ k_1 = f(t_n + \frac{1}{2}h, y_n + \frac{h}{2}k_1) \end{cases}$$

二级二阶公式：

$$\begin{cases} y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2) \\ k_1 = f(t_n, y_n) \\ k_2 = f\left(t_n + h, y_n + \frac{h}{2}(k_1 + k_2)\right) \end{cases}$$

二级四阶公式：

$$\begin{cases} y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2) \\ k_1 = f\left(t_n + \left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)h, y_n + \frac{h}{4}k_1 + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)hk_2\right) \\ k_2 = f\left(t_n + \left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)h, y_n + \left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)hk_1 + \frac{h}{4}k_2\right) \end{cases}$$

隐式 R - K 方法是 A-稳定的，并且可以构造出 N 级 $2N$ 阶隐式 R - K 方法。用 N 级隐式 R - K方法求解一个 s 维的微分方程组初值问题，每计算一步都要求解一个含 sN 个未知量 $k_{ri} (r=1,2,\cdots,s; i=1,2,\cdots,N)$的非线性方程组，因此，高级的隐式 R - K 方法在实际计算中不常用。

习　题

1. 试写出用 Euler 法求解初值问题

$$\begin{cases} y' = y(t-y), & 0 \leqslant t \leqslant 0.5 \\ y(0) = 1 \end{cases}$$

的计算公式，取步长$h=0.1$，并写出求解结果。

2. 试写出用 Euler 法求解初值问题

$$\begin{cases} y' = t-y, & 0 \leqslant t \leqslant 0.8 \\ y(0) = 2 \end{cases} \tag{7.67}$$

的计算公式，取步长 $h=0.2$，写出求解结果，并与精确解 $y(t)=3e^{-t}+t-1$ 作比较。

3. 试写出用改进的 Euler 法求解初值问题

$$\begin{cases} y' = \dfrac{2}{y-t}+1, & 0 \leqslant t \leqslant 1 \\ y(0) = 2 \end{cases}$$

的计算公式，取步长 $h=0.2$，并写出求解结果。

4. 试写出用中点公式(7.20)求解初值问题

$$\begin{cases} \dfrac{dy}{dt} = 2\sqrt{y-1}, & 0 \leqslant t \leqslant 1 \\ y(0) = 2 \end{cases}$$

的计算公式，取步长 $h=0.2$，并写出求解结果，再与精确解 $y(t)=1+(t+1)^2$ 作比较。

5. 试写出用四阶 Runge - Kutta 方法(7.27)求解初值问题(7.67)的计算公式，取步长 $h=0.2$，并写出计算结果，再与精确解作比较。

6. 试验证：求解初值问题(7.1)、(7.2)的单步法

$$y_{n+1} = y_n + \frac{h}{8}\left[3f(t_n,y_n) + 5f\left(t_n+\frac{4}{5}h, y_n+\frac{4}{5}hf(t_n,y_n)\right)\right]$$

是二阶方法。

7. 下列单步法是否与微分方程(7.1)相容：

(1) $y_{n+1} = y_n + \dfrac{h}{2}[f(t_n,y_n) + 2f(t_n+h, y_n+hf(t_n,y_n))]$；

(2) $y_{n+1} = y_n + hf\left(t_n+\dfrac{h}{2}, \dfrac{1}{2}y_n+\dfrac{h}{2}f(t_n,y_n)\right)$。

8. 设函数 $f(t,y)$在区域

$$D_0 = \{(t,y) \mid t_0 \leqslant t \leqslant T, |y| < \infty\}$$

内连续和有界且对变量 t 和 y 满足 Lipschitz 条件，试证：改进的 Euler 法是收敛的方法(设步长 h 的范围为 $0 \leqslant h \leqslant h_0$)。

9. 用 Euler 法求解下列初值问题时，由绝对稳定性对步长各有何限制：

(1) $\begin{cases} y' = -5y+\ln(t+1), & 0 \leqslant t \leqslant T \\ y(0) = y_0 \end{cases}$；　(7.68)

(2) $\begin{cases} y'=\mathrm{e}^{-t^2}-3y\sin t, & 0\leqslant t\leqslant\dfrac{\pi}{3} \\ y(0)=1 \end{cases}$ 。 (7.69)

10. 用二级二阶 R－K 方法求解初值问题

$$\begin{cases} y' = 1-\dfrac{2ty}{1+t^2}, & 0\leqslant t\leqslant 2 \\ y(0)=0 \end{cases}$$

时，由绝对稳定性对步长有何限制？

11. 证明：四阶 R－K 方法(7.27)的绝对稳定区域为

$$\left|1+h\lambda+\frac{(h\lambda)^2}{2}+\frac{(h\lambda)^3}{6}+\frac{(h\lambda)^4}{24}\right|<1$$

12. 用四阶 R－K 方法(7.27)求解初值问题(7.68)和(7.69)时，由绝对稳定性对步长 h 各有何限制？

13. 试确定系数 α 和 β，使线性二步法

$$y_{n+1} = \alpha y_{n-1} + \beta h f_n$$

具有尽可能高的阶，并求局部截断误差和阶数。

14. 试确定具有最高阶的线性三步显式方法，并求局部截断误差和阶数。

15. 试确定具有最高阶的线性三步法，并求局部截断误差和阶数。

16. 试确定三步三阶的 Gear 方法的系数，并求局部截断误差。

17. 证明：线性二步法

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1}+8f_n-f_{n-1})$$

与微分方程(7.1)相容，并且当计算开始值 y_1 的单步法是收敛的方法时它也是收敛的。

18. 证明：对于初值问题(7.1)、(7.2)，当计算开始值 y_1，y_2 的单步法是收敛的方法时，三步三阶的 Gear 方法也是收敛的。

19. 试检查下列线性多步法是否与微分方程(7.1)相容，是否满足根条件：

(1) $y_{n+1} = 2y_n + \dfrac{h}{12}(5f_{n+1}+8f_n-f_{n-1})$；

(2) $y_{n+2} = 2y_{n+1} - y_n + h(f_{n+1}-f_n)$；

(3) $y_{n+2} = y_n + h(3f_{n+1}-f_n)$。

20. 试求下列差分方程的通解：

$$e_{n+3}-7e_{n+2}+16e_{n+1}-12e_n=0 \quad (n=0,1,\cdots)$$

21. 试用边界轨迹法求下列线性多步法的绝对稳定区域：

(1) $y_{n+2} = y_n + \dfrac{h}{2}(f_{n+1}+3f_n)$； (7.70)

(2) Hamming 方法(7.41)。

22. 试求下列线性二步法的绝对稳定区间：

(1) $y_{n+2} = \dfrac{1}{2}(3y_{n+1}-y_n) + \dfrac{h}{24}(11f_{n+2}+8f_{n+1}-7f_n)$；

(2) $y_{n+1} = y_{n-1} + 2hf_n$；

(3) $y_{n+1} = \frac{1}{2}(y_n + y_{n-1}) + \frac{h}{4}(4f_{n+1} - f_n + 3f_{n-1})$。

23. 试写出用四阶 R－K 方法(7.27)求解初值问题

$$\begin{cases} y' = t + y - z, & y(0) = 1 \\ z' = tyz, & z(0) = 2 \end{cases} \quad 0 \leqslant t \leqslant 10$$

的计算公式(步长用 h 表示)。

24. 试写出用线性二步法(7.70)求解初值问题

$$\begin{cases} y'' = -200y - 20y', & 0 \leqslant t \leqslant 10 \\ y(0) = 1, & y'(0) = 1 \end{cases}$$

的计算公式，并判断：由绝对稳定性对步长 h 有何限制？

25. 试用四阶 R－K 法(7.27)求解初值问题

$$\begin{cases} y'' + \sin y = 0, & 0 \leqslant t \leqslant 2 \\ y(0) = 1, & y'(0) = 0 \end{cases}$$

(取步长 $h=0.4$)。

26. 用改进的 Euler 法和 $k=6$ 的 Gear 方法求解初值问题

$$\begin{cases} x' = y - 7x + \sqrt{t}, & x(0) = 1 \\ y' = -2x - 5y + t^2, & y(0) = 1 \end{cases} \quad 0 \leqslant t \leqslant 10$$

由绝对稳定性对步长 h 各有何限制？

27. 试写出用中点公式(7.20)求解初值问题

$$\begin{cases} x' = -2x + ty - z\cos t, & x(0) = 0 \\ y' = -2y + z\sin t + t, & y(0) = 1 \\ z' = y\sin t - 2z - t^2, & z(0) = 2 \end{cases} \quad 0 \leqslant t \leqslant 10$$

的计算公式，并判断：由绝对稳定性对步长 h 有何限制？

28. 用 Euler 法求解初值问题

$$\begin{cases} x' = -x + ty + \varphi_1(t), & x(0) = x_0 \\ y' = -x - 2y + \varphi_2(t), & y(0) = y_0 \end{cases} \quad 0 \leqslant t \leqslant 1$$

时，由绝对稳定性对步长 h 有何限制？

29. 求下列方程组的刚性比

$$\begin{cases} x' = -10x + 9y, & x(t_0) = x_0 \\ y' = 10x - 11y, & y(t_0) = y_0 \end{cases} \quad t_0 \leqslant t \leqslant T$$

又问：用四阶 R－K 方法(7.27)求解时，由绝对稳定性对步长 h 有何限制？

30. 设有微分方程组

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{\Phi}(t)$$

其中 $\boldsymbol{A}$ 是 $s \times s$ 的常数矩阵，其特征值 $\lambda_i (i=1,2,\cdots,s)$ 均为负实数，且刚性比是 10^8。如果用 Euler 法求解，并从 $t=0$ 计算到满足条件

$$e^{\lambda_i t} < 0.001 \quad (i = 1,2,\cdots,s)$$

的 t，问：至少要计算多少步？

第 8 章　偏微分方程的差分解法

本章只简要介绍几种最典型的偏微分方程定解问题的差分解法。

8.1　椭圆型方程第一边值问题

椭圆型方程最典型的是 Poisson(泊松)方程

$$\Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y) \tag{8.1}$$

若 $f(x,y) \equiv 0$，则方程(8.1)成为

$$\Delta u = 0 \tag{8.2}$$

方程(8.2)称为 Laplace(拉普拉斯)方程。椭圆型方程的定解条件主要是边界条件。

定解问题

$$\begin{cases} \Delta u = f(x,y), & (x,y) \in \Omega \quad (8.3) \\ u|_{\Gamma} = \varphi(x,y), & (x,y) \in \Gamma \quad (8.4) \end{cases}$$

称为 Poisson 方程第一边值问题。其中 Ω 是 Oxy 坐标面上的一个有界区域，Γ 是 Ω 的边界，条件(8.4)称为边界条件，$f(x,y)$ 和 $\varphi(x,y)$ 都是已知函数。求解第一边值问题(8.3)、(8.4)就是要求出未知函数 $u(x,y)$，使它在区域 Ω 内满足方程(8.3)，在 Ω 的边界 Γ 上满足边界条件(8.4)。总假定问题(8.3)、(8.4)是适定的，即它的解 $u(x,y)$ 存在、唯一且连续地依赖于函数 $f(x,y)$ 和 $\varphi(x,y)$，又假定它的解 $u(x,y)$ 在 Ω 内充分光滑。

差分解法和有限元法是求解偏微分方程定解问题的两种主要数值解法，但这里只介绍差分解法。本节将介绍第一边值问题(8.3)、(8.4)的差分解法。

8.1.1　差分方程的建立

在 Oxy 坐标面上取定一点 (x_0, y_0)，用两组平行直线

$$x = x_0 + ih \quad (i = 0, \pm 1, \pm 2, \cdots)$$
$$y = y_0 + j\tau \quad (j = 0, \pm 1, \pm 2, \cdots)$$

构成矩形网格覆盖整个 Oxy 坐标面，见图 8－1，$h>0$，$\tau>0$ 分别是 x 方向和 y 方向的步长。两组平行线的交点 (x_i, y_j) $(i,j=0,\pm 1,\pm 2,\cdots)$ 称为节点。如果两个节点沿 x 方向(或 y 方向)的距离只差一个步长，则称为相邻节点。若一个节点属于 $\Omega \cup \Gamma$，并且它的四个相邻节点均属于 $\Omega \cup \Gamma$，则此节点称为正则内节点(图8－1中画○者)，正则内节点的集合记为 $\Omega'_{h,\tau}$；若一个节点属于 $\Omega \cup \Gamma$，并且它的四个相邻节点中至少有一个不属于 $\Omega \cup \Gamma$，则此节点称为非正则内节点(图 8－1 中画●者)。网格线与边界线 Γ 的交点称为边界点(图 8－1 中画×者)。问题是要求出第一边值问题(8.3)、(8.4)在正则与非正则内节点上的数值解。

为简化记号，把节点 (x_i, y_j) 简记为 (i,j)，把 $u(x_i, y_j)$ 简记为 $u(i,j)$。由一元函数的 Taylor 级数，可得二阶偏导数的下列表达式

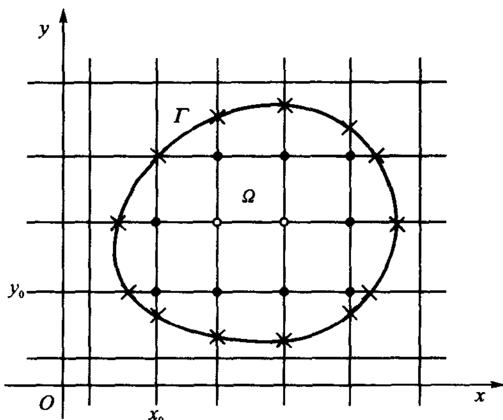


图 8-1　正则内节点、非正则内节点与边界点

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{(i,j)} = \frac{1}{h^2}[u(i+1,j) - 2u(i,j) + u(i-1,j)] + O(h^2)$$

$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{(i,j)} = \frac{1}{\tau^2}[u(i,j+1) - 2u(i,j) + u(i,j-1)] + O(\tau^2)$$

在正则内节点$(i,j)\in\Omega'_{h,\tau}$处，偏微分方程(8.3)可表达为

$$\begin{aligned}\Delta_{h,\tau}u(i,j) &\equiv \frac{1}{h^2}[u(i+1,j) - 2u(i,j) + u(i-1,j)] + \\ &\frac{1}{\tau^2}[u(i,j+1) - 2u(i,j) + u(i,j-1)] = \\ &f_{i,j} + O(h^2) + O(\tau^2)\end{aligned} \tag{8.5}$$

其中 $f_{i,j}=f(x_i,y_j)$。在式(8.5)中略去 $O(h^2)+O(\tau^2)$，得到偏微分方程(8.3)在点$(i,j)\in\Omega'_{h,\tau}$处的近似方程

$$\Delta_{h,\tau}u(i,j) \approx f_{i,j} \tag{8.6}$$

把此近似方程的近似等号改为等号，又用 $u_{i,j}$表示 $u(i,j)$的近似值，就得到偏微分方程(8.3)在正则内节点处的差分近似

$$\begin{aligned}\Delta_{h,\tau}u_{i,j} &\equiv \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + \\ &\frac{1}{\tau^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = f_{i,j}, \quad (i,j)\in\Omega'_{h,\tau}\end{aligned} \tag{8.7}$$

当 $u(i,j)$是偏微分方程(8.3)的解时，近似等式(8.6)的误差为 $O(h^2)+O(\tau^2)$，简记为 $O(h^2+\tau^2)$。这个误差称为用差分方程(8.7)逼近偏微分方程(8.3)的截断误差，并且称差分方程(8.7)对 x 和 y 都是二阶精度的。

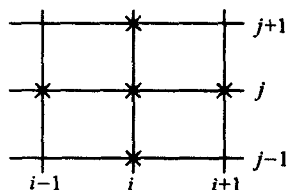


图 8-2　五点格式的节点

差分方程又称为差分格式。由于式(8.7)用到五个节点(见图 8-2)，故又称式(8.7)为五点格式。当 $h=\tau$ 时，差分方程(8.7)简化为

$$\Delta_h u_{i,j} \equiv \frac{1}{h^2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}) = f_{i,j}, \quad (i,j) \in \Omega'_h \tag{8.8}$$

这里 Ω'_h 表示当 $h=\tau$ 时正则内节点的集合。

在差分方程(8.7)或者(8.8)中,方程的个数等于正则内节点的数目,但未知数除了正则内节点上的 u 值之外,还包括一些非正则内节点上的 u 值。所以,要计算出微分方程(8.3)在所有内节点上的数值解,单靠方程(8.7)或(8.8)是不够的,必须使用边界条件。

8.1.2 边界条件的使用

这里仅讨论 $h=\tau$ 的情形。

若非正则内节点恰好是边界点,图 8-3 中的点 A 就是这种点,则由边界条件(8.4)可知

$$u_A = \varphi(A)$$

对于不是边界点的非正则内节点,例如图 8-3 中的 C 点,有两种处理方法。

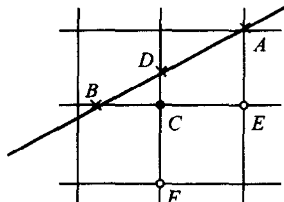


图 8-3 非正则内节点 C 与边界点的关系

一种是直接转移法。取与点 C 距离较近的边界点(例如图 8-3 中的点 D)上的 u 值作为 $u(C)$ 的近似值 u_C,即

$$u_C = u(D) = \varphi(D) \approx u(C) \tag{8.9}$$

这时有

$$|u(C) - u_C| = |u'_y(x_D, \xi)| \, |CD| < h \, |u'_y(x_D, \xi)|, \quad y_C < \xi < y_D$$

可见,近似式(8.9)的截断误差为 $O(h)$。

另一种是线性插值法。在图 8-3 中,取边界点 B 与正则内节点 E(也可取边界点 D 与正则内节点 F)作为插值节点对点 C 作线性插值,得

$$u_C = \frac{x_C - x_E}{x_B - x_E}u(B) + \frac{x_C - x_B}{x_E - x_B}u(E) = \frac{h}{h+\delta}\varphi(B) + \frac{\delta}{h+\delta}u(E) \tag{8.10}$$

其中 $\delta = |BC|$。根据线性插值公式的余项结构,有

$$|u(C) - u_C| = \left|\frac{1}{2} \cdot \frac{\partial^2 u(\xi, y_C)}{\partial x^2}(x_C - x_B)(x_C - x_E)\right| <$$

$$\frac{h^2}{2}\left|\frac{\partial^2 u(\xi, y_C)}{\partial x^2}\right|, \quad x_B < \xi < x_E$$

可见用式(8.10)右端表示 $u(C)$ 的近似值 u_C,其截断误差是 $O(h^2)$,与差分方程(8.8)的截断误差同阶。

总之,利用边界条件(8.4)对非正则内节点进行以上的处理所得到的方程与式(8.8)联立,就组成了一个方程数目与未知数数目都等于全部内节点数的线性代数方程组。求解这个方程组就得到第一边值问题(8.3)、(8.4)的数值解。

例 1 设 $\Omega \cup \Gamma$ 如图 8-4 所示。网格为正方形,步长为 h。试写出第一边值问题(8.3)、(8.4)的差分格式。

解 由式(8.8),得微分方程在正则内节点上的差分近似

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 f_{i,j}$$
$$(i,j) = (1,1),(2,1)$$

由边界条件得

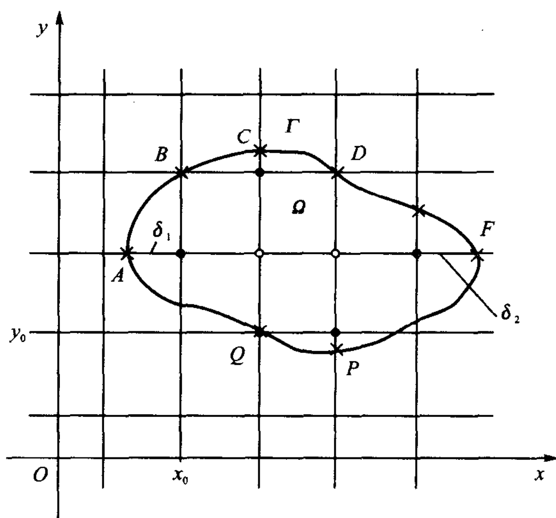


图 8－4　例 1 附图

$$u_{0,2}=\varphi(B),\quad u_{2,2}=\varphi(D),\quad u_{1,0}=\varphi(Q)$$

又用直接转移法，得

$$u_{1,2}=\varphi(C),\quad u_{2,0}=\varphi(P)$$

用线性插值法，分别得到关于非正则内节点(0,1)和(3,1)的方程

$$u_{0,1}=\frac{h}{h+\delta_1}\varphi(A)+\frac{\delta_1}{h+\delta_1}u_{1,1}$$

$$u_{3,1}=\frac{\delta_2}{h+\delta_2}u_{2,1}+\frac{h}{h+\delta_2}\varphi(F)$$

可见，共有四个未知数 $u_{0,1}, u_{1,1}, u_{2,1}, u_{3,1}$ 和四个方程。未知数的排列次序通常按照所在节点从下到上、从左到右的原则；各个方程则按照它们所对应的节点的次序排列。于是，所求的差分格式为

$$\begin{cases} u_{0,1}-\dfrac{\delta_1}{h+\delta_1}u_{1,1}=\dfrac{h}{h+\delta_1}\varphi(A) \\ u_{0,1}-4u_{1,1}+u_{2,1}=h^2f_{1,1}-\varphi(C)-\varphi(Q) \\ u_{1,1}-4u_{2,1}+u_{3,1}=h^2f_{2,1}-\varphi(D)-\varphi(P) \\ -\dfrac{\delta_2}{h+\delta_2}u_{2,1}+u_{3,1}=\dfrac{h}{h+\delta_2}\varphi(F) \end{cases}$$

它的矩阵形式为

$$\begin{bmatrix} 1 & -\dfrac{\delta_1}{h+\delta_1} & & \\ 1 & -4 & 1 & \\ & 1 & -4 & 1 \\ & & -\dfrac{\delta_2}{h+\delta_2} & 1 \end{bmatrix}\begin{bmatrix} u_{0,1} \\ u_{1,1} \\ u_{2,1} \\ u_{3,1} \end{bmatrix}=\begin{bmatrix} \dfrac{h}{h+\delta_1}\varphi(A) \\ h^2f_{1,1}-\varphi(C)-\varphi(Q) \\ h^2f_{2,1}-\varphi(D)-\varphi(P) \\ \dfrac{h}{h+\delta_2}\varphi(F) \end{bmatrix}$$

8.1.3 差分方程组解的存在唯一性

设网格是正方形的，边界条件的处理采用直接转移法，则相应于第一边值问题(8.3)、(8.4)的差分方程组为

$$
\begin{cases}
\Delta_h u_{i,j} \equiv \dfrac{1}{h^2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}) = f_{i,j}, & (i,j) \in \Omega'_h \\
u_{i,j} = \varphi(P), \quad (i,j) \in \Omega''_h, P \in \Gamma_h
\end{cases} \tag{8.11}
$$

其中 P 是 Γ_h 中最靠近 $(i,j)\in\Omega''_h$ 的点，Ω''_h 是非正则内节点集合，Γ_h 是边界点集合。式(8.11)又称为差分方程边值问题。

定理 8.1(极值原理) 设 $\{v_{i,j}\}$ 是定义在点集 $\Omega'_h \cup \Omega''_h$ 上的一组数，如果

(1) $v_{i,j} \not\equiv$ 常数，$(i,j)\in\Omega'_h\cup\Omega''_h$；

(2) $\Delta_h v_{i,j} \geqslant 0(\Delta_h v_{i,j} \leqslant 0)$，$(i,j)\in\Omega'_h$。

则数组 $\{v_{i,j}\}$ 不可能在点集 Ω'_h 上取得正的最大值(负的最小值)。

证 用反证法。设存在节点 $(i_0,j_0)\in\Omega'_h$，使

$$v_{i_0,j_0} = \max\{v_{i,j}\} > 0$$

由条件(1)可知，在与 (i_0,j_0) 相邻的节点中，至少有一个节点 (i_1,j_1)，使

$$v_{i_1,j_1} < v_{i_0,j_0}$$

因而有

$$\Delta_h v_{i_0,j_0} < 0$$

这与条件(2)相矛盾。因此，$\{v_{i,j}\}$ 不可能在点集 Ω'_h 上取得正的最大值。

同理可证另一种情况。

证毕。

定理 8.2 差分方程组(8.11)的解存在且唯一。

证 只须证明相应的齐次线性方程组

$$
\begin{cases}
\Delta_h u_{i,j} = 0, \quad (i,j) \in \Omega'_h & (8.12) \\
u_{i,j} = 0, \quad (i,j) \in \Omega''_h & (8.13)
\end{cases}
$$

只有零解。

若 $u_{i,j} \equiv c$(常数)，$(i,j)\in\Omega'_h\cup\Omega''_h$，则由式(8.13)知 $c=0$。

今设 $u_{i,j} \not\equiv c$(常数)，$(i,j)\in\Omega'_h\cup\Omega''_h$，则由式(8.12)并根据极值原理的第一种情况可知 $\{u_{i,j}\}$ 只能在 Ω''_h 上取得正的最大值。但在 Ω''_h 上 $u_{i,j}=0$，故有

$$u_{i,j} \leqslant 0, \quad (i,j) \in \Omega'_h \cup \Omega''_h$$

又由式(8.12)并根据极值原理的第二种情况可知 $\{u_{i,j}\}$ 只能在 Ω''_h 上取得负的最小值，但在 Ω''_h 上 $u_{i,j}=0$，故有

$$u_{i,j} \geqslant 0, \quad (i,j) \in \Omega'_h \cup \Omega''_h$$

综合两者，可知

$$u_{i,j} = 0, \quad (i,j) \in \Omega'_h \cup \Omega''_h$$

证毕。

8.2 抛物型方程初边值问题

抛物型方程最典型的是常系数扩散方程

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} \tag{8.14}$$

其中常数 $a>0$。常系数扩散方程的定解问题分为初值问题和初边值问题两种。初值问题是

$$\begin{cases} \dfrac{\partial u}{\partial t} = a\dfrac{\partial^2 u}{\partial x^2}, & 0<t<T, \quad -\infty<x<\infty \quad (8.15) \\ u(x,0) = \varphi(x), & -\infty<x<\infty \quad (8.16) \end{cases}$$

其中 $\varphi(x)$是已知函数，式(8.16)称为初始条件。初边值问题为

$$\begin{cases} \dfrac{\partial u}{\partial t} = a\dfrac{\partial^2 u}{\partial x^2}, & 0<t<T, \quad 0<x<l \quad (8.17) \\ u(x,0) = \varphi(x), & 0\leqslant x\leqslant l \quad (8.18) \\ u(0,t) = \mu_1(t), & u(l,t) = \mu_2(t), \quad 0\leqslant t\leqslant T \quad (8.19) \end{cases}$$

其中 $\mu_1(t),\mu_2(t),\varphi(x)$是已知函数，并且满足 $\varphi(0)=\mu_1(0),\varphi(l)=\mu_2(0)$。式(8.19)称为第一类边界条件。边界条件也可以是

$$\begin{cases} \left(\dfrac{\partial u}{\partial x} - \lambda_1(t)u\right)\Bigg|_{x=0} = \mu_1(t) \\ \left(\dfrac{\partial u}{\partial x} + \lambda_2(t)u\right)\Bigg|_{x=l} = \mu_2(t) \end{cases} \quad 0\leqslant t\leqslant T \tag{8.20}$$

其中 $\lambda_1(t)\geqslant0,\lambda_2(t)\geqslant0,\mu_1(t),\mu_2(t)$都是已知函数。当 $\lambda_1(t)$和 $\lambda_2(t)$都恒为零时，式(8.20)称为第二类边界条件；当$|\lambda_1(t)|+|\lambda_2(t)|\not\equiv0$ 时，式(8.20)称为第三类边界条件。

总假定上述定解问题都是适定的，并且它们的解 $u(x,t)$在求解区域内充分光滑。

定解问题(8.15)、(8.16)的求解区域是

$$\Omega_1: \quad 0\leqslant t\leqslant T, \quad -\infty<x<\infty$$

而定解问题(8.17)～(8.19)的求解区域是

$$\Omega_2: \quad 0\leqslant t\leqslant T, \quad 0\leqslant x\leqslant l$$

在这两个定解问题中，t 一般表示时间，称为时间坐标；x 一般表示位置，称为空间坐标。

8.2.1　差分方程的建立与定解条件的离散化

用差分解法求解上述定解问题，首先要对求解区域网格化。用平行于 x 轴和平行于 t 轴的直线将求解区域分为若干个长方形网格，x 方向的步长记为 h，t 方向的步长记为 τ，h 和 τ 皆为正的常数。网格节点称为节点。对于区域 Ω_1，节点坐标是

$$\begin{cases} x_k = kh & (k=0,\pm1,\pm2,\cdots) \\ t_j = j\tau & \left(j=0,1,\cdots,m;\ m=\left[\dfrac{T}{\tau}\right]\right) \end{cases}$$

对于区域 Ω_2，节点坐标是

$$\begin{cases} x_k = kh & (k=0,1,\cdots,N) \quad h=\dfrac{l}{N} \\ t_j = j\tau & \left(j=0,1,\cdots,m;\ m=\left[\dfrac{T}{\tau}\right]\right) \end{cases}$$

参看图 8－5 和图 8－6。位于求解区域边界上的节点称为边界节点，其余位于求解区域内部的节点称为内节点。

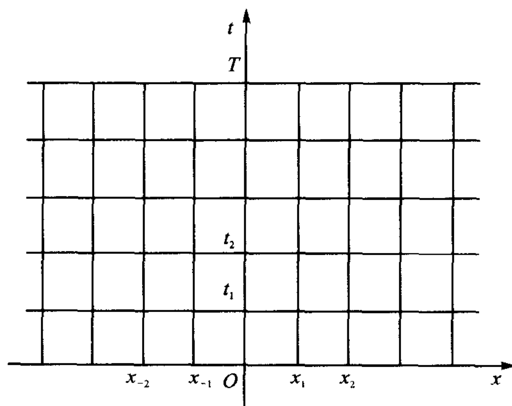


图 8-5 初值问题求解区域的网格化

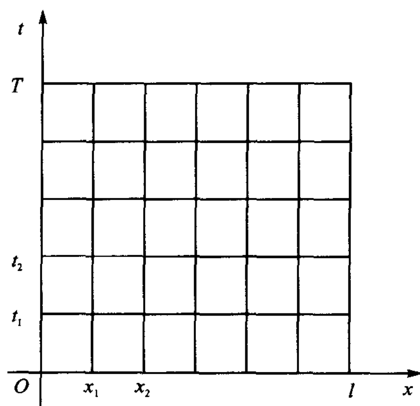


图 8-6 初边值问题求解区域的网格化

类似于 8.1 节的规定，点(x_k, t_j)简记为(k,j)，函数值$u(x_k,t_j)$简记为$u(k,j)$，用$u_{k,j}$表示$u(k,j)$的近似值。

利用一元函数的 Taylor 级数，可推出下列偏导数的差商表达式：

$$\left(\frac{\partial u}{\partial t}\right)_{(k,j)} = \frac{u(k,j+1)-u(k,j)}{\tau} + O(\tau) \tag{8.21}$$

$$\left(\frac{\partial u}{\partial t}\right)_{(k,j)} = \frac{u(k,j)-u(k,j-1)}{\tau} + O(\tau) \tag{8.22}$$

$$\left(\frac{\partial u}{\partial t}\right)_{(k,j)} = \frac{u(k,j+1)-u(k,j-1)}{2\tau} + O(\tau^2) \tag{8.23}$$

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{(k,j)} = \frac{u(k+1,j)-2u(k,j)+u(k-1,j)}{h^2} + O(h^2) \tag{8.24}$$

1. 古典显式格式

利用公式(8.21)和(8.24)，在内节点(k,j)处，偏微分方程(8.14)可表示为

$$\frac{u(k,j+1)-u(k,j)}{\tau} - a\frac{u(k+1,j)-2u(k,j)+u(k-1,j)}{h^2} = O(\tau+h^2)$$

这里，$O(\tau+h^2)$表示$O(\tau)+O(h^2)$。在上述等式中略去$O(\tau+h^2)$项，得到偏微分方程(8.14)在内节点(k,j)处的近似方程

$$\frac{u(k,j+1)-u(k,j)}{\tau} \approx a\frac{u(k+1,j)-2u(k,j)+u(k-1,j)}{h^2} \tag{8.25}$$

把式(8.25)中的近似等号改为等号，就得到在内节点(k,j)处逼近偏微分方程(8.14)的差分方程

$$\frac{u_{k,j+1}-u_{k,j}}{\tau} = a\frac{u_{k+1,j}-2u_{k,j}+u_{k-1,j}}{h^2}$$

为便于计算，把此差分方程改写成

$$u_{k,j+1} = aru_{k-1,j} + (1-2ar)u_{k,j} + aru_{k+1,j} \tag{8.26}$$

其中$r=\frac{\tau}{h^2}$，称r为网格比。

当 u 是方程(8.14)的解时，近似等式(8.25)的误差 $O(\tau+h^2)$ 称为差分方程(8.26)逼近偏微分方程(8.14)的截断误差。可见，差分方程(8.26)对时间 t 是一阶精度的，对空间 x 是二阶精度的。

差分方程(8.26)所用到的节点如图 8-7 所示。出现在式(8.26)中的节点是相邻两个时间层上的节点，并且在后一时间层上只含一个节点处的 u 值 $u_{k,j+1}$，只要知道前一时间层上三个节点处的 u 值 $u_{k-1,j}$，$u_{k,j}$ 和 $u_{k+1,j}$ 就可由式(8.26)直接计算 $u_{k,j-1}$。称差分方程(8.26)为二层显式格式，又称为古典显式格式。

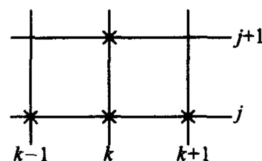


图 8-7　古典显式格式的节点

为了把方程(8.14)化为差分方程求解，除了对方程(8.14)建立差分近似外，还要对定解条件进行离散化。对初始条件(8.16)和(8.18)分别离散化为

$$u_{k,0}=\varphi(kh)\quad(k=0,\pm1,\pm2,\cdots)\tag{8.27}$$

$$u_{k,0}=\varphi(kh)\quad(k=1,2,\cdots,N-1)\tag{8.28}$$

对于第一类边界条件(8.19)，离散化为

$$u_{0,j}=\mu_1(j\tau),\quad u_{N,j}=\mu_2(j\tau)\quad(j=0,1,\cdots,m)\tag{8.29}$$

对于第二、三类边界条件(8.20)，使用公式

$$\left(\frac{\partial u}{\partial x}\right)_{(0,j)}=\frac{u(1,j)-u(0,j)}{h}+O(h)$$

$$\left(\frac{\partial u}{\partial x}\right)_{(N,j)}=\frac{u(N,j)-u(N-1,j)}{h}+O(h)$$

得到式(8.20)在边界节点 $(0,j)$ 和 (N,j) 处的差分近似

$$\begin{cases}\dfrac{u_{1,j}-u_{0,j}}{h}-\lambda_1(j\tau)u_{0,j}=\mu_1(j\tau)\\[2mm]\dfrac{u_{N,j}-u_{N-1,j}}{h}+\lambda_2(j\tau)u_{N,j}=\mu_2(j\tau)\\[2mm]\qquad(j=0,1,\cdots,m)\end{cases}\tag{8.30}$$

这种差分近似的截断误差是 $O(h)$，其精度是一阶的，比差分方程(8.26)对空间 x 的精度低一阶。为提高(8.20)离散化后的精度，改用公式

$$\begin{cases}\left(\dfrac{\partial u}{\partial x}\right)_{(0,j)}=\dfrac{u(1,j)-u(-1,j)}{2h}+O(h^2)\\[2mm]\left(\dfrac{\partial u}{\partial x}\right)_{(N,j)}=\dfrac{u(N+1,j)-u(N-1,j)}{2h}+O(h^2)\end{cases}$$

得到式(8.20)在边界节点 $(0,j)$ 和 (N,j) 处的另一种差分近似

$$\begin{cases}\dfrac{u_{1,j}-u_{-1,j}}{2h}-\lambda_1(j\tau)u_{0,j}=\mu_1(j\tau)\\[2mm]\dfrac{u_{N+1,j}-u_{N-1,j}}{2h}+\lambda_2(j\tau)u_{N,j}=\mu_2(j\tau)\end{cases}\tag{8.31}$$

这种差分近似的截断误差是 $O(h^2)$。但是式(8.31)中含有求解区域之外的点 $(-1,j)$ 和 $(N+1,j)$。因此必须设法消去式(8.31)中的 $u_{-1,j}$ 和 $u_{N+1,j}$。

假定扩散方程(8.17)在边界上也成立，则可以把内节点处的差分方程(8.26)推广到边界

节点上。在左边界节点上，有

$$u_{0,j+1} = aru_{-1,j} + (1-2ar)u_{0,j} + aru_{1,j} \tag{8.32}$$

由式(8.31)的第一个方程与式(8.32)联立消去 $u_{-1,j}$，得到

$$u_{0,j+1} = \{1 - 2ar[1 + h\lambda_1(j\tau)]\}u_{0,j} + 2aru_{1,j} - 2arh\mu_1(j\tau) \tag{8.33}$$

在右边界节点上，式(8.26)成为

$$u_{N,j+1} = aru_{N-1,j} + (1-2ar)u_{N,j} + aru_{N+1,j} \tag{8.34}$$

由式(8.31)的第二个方程与式(8.33)联立消去 $u_{N-1,j}$，得到

$$u_{N,j+1} = 2aru_{N-1,j} + \{1 - 2ar[1 + h\lambda_2(j\tau)]\}u_{N,j} + 2arh\mu_2(j\tau) \tag{8.35}$$

公式(8.33)和(8.35)给出了第二、三类边界条件(8.20)的差分近似，其截断误差是 $O(h^2)$。

偏微分方程的差分近似与离散化了的定解条件一起构成了一个差分方程定解问题，用它的解近似代替原偏微分方程定解问题的解。

对于初值问题(8.15)、(8.16)，使用式(8.26)和(8.27)，得到如下的差分方程定解问题

$$\begin{cases} u_{k,j+1} = aru_{k-1,j} + (1-2ar)u_{k,j} + aru_{k+1,j} \\ \qquad (k = 0, \pm 1, \pm 2, \cdots; j = 0, 1, \cdots, m-1) \\ u_{k,0} = \varphi(kh) \quad (k = 0, \pm 1, \pm 2, \cdots) \end{cases} \tag{8.36}$$

当初始条件给定后，由格式(8.36)可逐层逐点计算节点上的 u 值。

对于第一类初边值问题(8.17)、(8.18)和(8.19)，使用式(8.26)、(8.28)、(8.29)，得到

$$\begin{cases} u_{k,j+1} = aru_{k-1,j} + (1-2ar)u_{k,j} + aru_{k+1,j} \\ \qquad (k = 1, 2, \cdots, N-1; j = 0, 1, \cdots, m-1) \\ u_{k,0} = \varphi(kh) \quad (k = 1, 2, \cdots, N-1) \\ u_{0,j} = \mu_1(j\tau), \quad u_{N,j} = \mu_2(j\tau) \quad (j = 0, 1, \cdots, m) \end{cases} \tag{8.37}$$

格式(8.37)也是逐层逐点计算节点上的 u 值。格式(8.37)还可表示成矩阵形式

$$\begin{cases} \boldsymbol{u}_{j+1} = \boldsymbol{A}\boldsymbol{u}_j + \boldsymbol{f}_j \quad (j = 0, 1, \cdots, m-1) \\ \boldsymbol{u}_0 = \boldsymbol{\Phi} \end{cases}$$

其中

$$\boldsymbol{A} = \begin{bmatrix} 1-2ar & ar & & & \\ ar & 1-2ar & ar & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & ar \\ & & & ar & 1-2ar \end{bmatrix}$$

$$\boldsymbol{u}_j = (u_{1,j}, u_{2,j}, \cdots, u_{N-1,j})^{\mathrm{T}}$$

$$\boldsymbol{f}_j = (ar\mu_1(j\tau), 0, \cdots, 0, ar\mu_2(j\tau))^{\mathrm{T}}$$

$$\boldsymbol{\Phi} = (\varphi(h), \varphi(2h), \cdots, \varphi((N-1)h))^{\mathrm{T}}$$

对第二、三类初边值问题(8.17)、(8.18)、(8.20)，使用式(8.26)、(8.28)、(8.33)、(8.35)，得到

$$
\begin{cases}
u_{k,j+1} = aru_{k-1,j} + (1-2ar)u_{k,j} + aru_{k+1,j} \\
\qquad (k = 1,2,\cdots,N-1) \\
u_{0,j+1} = \alpha_j u_{0,j} + 2aru_{1,j} - 2arh\mu_1(j\tau) \\
u_{N,j+1} = 2aru_{N-1,j} + \beta_j u_{N,j} + 2arh\mu_2(j\tau) \\
\qquad (j = 0,1,\cdots,m-1) \\
u_{k,0} = \varphi(kh) \quad (k = 0,1,\cdots,N)
\end{cases}
\tag{8.38}
$$

其中

$$
\alpha_j = 1 - 2ar[1 + h\lambda_1(j\tau)]
$$
$$
\beta_j = 1 - 2ar[1 + h\lambda_2(j\tau)]
$$

2. 古典隐式格式

利用公式(8.22)和(8.24)，仿照前述的方法，可得到在内节点(k,j)处逼近偏微分方程(8.14)的又一个差分方程

$$
\frac{u_{k,j} - u_{k,j-1}}{\tau} = a\,\frac{u_{k-1,j} - 2u_{k,j} + u_{k+1,j}}{h^2}
$$

或写成

$$
-aru_{k-1,j} + (1+2ar)u_{k,j} - aru_{k+1,j} = u_{k,j-1} \tag{8.39}
$$

它的截断误差为$O(\tau+h^2)$，其精度与古典显式格式(8.26)相同。在式(8.39)中用到的节点如图 8-8 所示。出现在后一时间层上的节点有三个。如果$u_{k,j-1}$已知，并不能单靠式(8.39)一个方程求出$u_{k-1,j}$，$u_{k,j}$和$u_{k+1,j}$。称差分方程(8.39)为二层隐式格式。这种格式的差分方程适用于求解初边值问题。

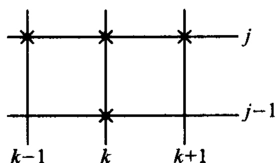


图 8-8　古典隐式格式的节点

对于第一类初边值问题(8.17)、(8.18)、(8.19)，使用差分方程(8.39)和离散化了的定解条件(8.28)、(8.29)，得到

$$
\begin{cases}
-aru_{k-1,j} + (1+2ar)u_{k,j} - aru_{k+1,j} = u_{k,j-1} \\
\qquad (k = 1,2,\cdots,N-1;\ j = 1,2,\cdots,m) \\
u_{k,0} = \varphi(kh) \quad (k = 1,2,\cdots,N-1) \\
u_{0,j} = \mu_1(j\tau),\quad u_{N,j} = \mu_2(j\tau) \quad (j = 0,1,\cdots,m)
\end{cases}
\tag{8.40}
$$

格式(8.40)的矩阵形式为

$$
\begin{cases}
\boldsymbol{B}\,\boldsymbol{u}_j = \boldsymbol{u}_{j-1} + \boldsymbol{f}_j \quad (j = 1,2,\cdots,m) \\
\boldsymbol{u}_0 = \boldsymbol{\phi}
\end{cases}
\tag{8.41}
$$

这里的矩阵 $\boldsymbol{B}$ 为

$$
\boldsymbol{B} = \begin{bmatrix}
1+2ar & -ar & & & \\
-ar & 1+2ar & -ar & & \\
 & \ddots & \ddots & \ddots & \\
 & & \ddots & \ddots & -ar \\
 & & & -ar & 1+2ar
\end{bmatrix}
$$

使用格式(8.41)求解时，每计算一层节点上的 u 值 $u_{1,j}$，$u_{2,j}$，$\cdots$，$u_{N-1,j}$，都要解一个线性代数

方程组。容易看出，它的系数矩阵 $\boldsymbol{B}$ 是主对角线元素按行严格占优阵，因此，对每一个 j，方程组(8.41)的解存在且唯一。

3. Crank - Nicolson 格式

利用 Taylor 级数容易证明下列两个等式成立：

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{(k,j+\frac{1}{2})} = \frac{1}{2}\left[\left(\frac{\partial^2 u}{\partial x^2}\right)_{(k,j+1)} + \left(\frac{\partial^2 u}{\partial x^2}\right)_{(k,j)}\right] + O(\tau^2)$$

$$\left(\frac{\partial u}{\partial t}\right)_{(k,j+\frac{1}{2})} = \frac{u(k,j+1) - u(k,j)}{\tau} + O(\tau^2)$$

利用这两个公式和公式(8.24)，可得偏微分方程(8.17)在点$\left(k,j+\dfrac{1}{2}\right)$处的差分近似

$$-aru_{k-1,j+1} + 2(1+ar)u_{k,j+1} - aru_{k+1,j+1} = aru_{k-1,j} + 2(1-ar)u_{k,j} + aru_{k+1,j} \tag{8.42}$$

其截断误差为 $O(\tau^2+h^2)$。因此，差分方程(8.42)逼近偏微分方程(8.17)的精度是二阶的。称(8.42)为 Crank - Nicolson(克兰克-尼科尔松)格式，它所用到的节点如图 8 - 9 所示，这种格式属于二层隐式格式。如果使用它求解第一类初边值问题，那么，由式(8.42)、(8.28)、(8.29)构成了如下的差分方程定解问题：

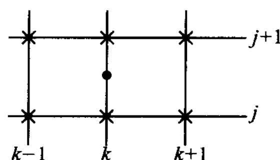


图 8 - 9 Crank - Nicolson 格式的节点

$$\begin{cases} -aru_{k-1,j+1} + 2(1+ar)u_{k,j+1} - aru_{k+1,j+1} = \\ \qquad aru_{k-1,j} + 2(1-ar)u_{k,j} + aru_{k+1,j} \\ \qquad\qquad (k = 1,2,\cdots,N-1; j = 0,1,\cdots,m-1) \\ u_{k,0} = \varphi(kh) \quad (k = 1,2,\cdots,N-1) \\ u_{0,j} = \mu_1(j\tau), \quad u_{N,j} = \mu_2(j\tau) \quad (j = 0,1,\cdots,m) \end{cases} \tag{8.43}$$

它的矩阵形式是

$$\begin{cases} (\boldsymbol{I}+\boldsymbol{B})\boldsymbol{u}_{j+1} = (\boldsymbol{I}+\boldsymbol{A})\boldsymbol{u}_j + \boldsymbol{f}_{j-1} + \boldsymbol{f}_j \\ \qquad (j = 0,1,\cdots,m-1) \\ \boldsymbol{u}_0 = \boldsymbol{\phi} \end{cases} \tag{8.44}$$

其中 $\boldsymbol{I}$ 是$(N-1)\times(N-1)$的单位矩阵。由于 $\boldsymbol{I}+\boldsymbol{B}$ 是主对角线元素按行严格占优阵，所以对每一个 j，方程组(8.44)的解存在且唯一。

如果使用格式(8.42)求解第二、三类初边值问题，并对边界条件(8.20)采用差分近似式(8.30)，则相应的差分方程定解问题为

$$\begin{cases} \boldsymbol{P}\boldsymbol{u}_{j+1} = \boldsymbol{Q}\boldsymbol{u}_j + \boldsymbol{g}_{j+1} \quad (j = 0,1,\cdots,m-1) \\ \boldsymbol{u}_0 = (\varphi(0),\varphi(h),\cdots,\varphi(Nh))^{\mathrm{T}} \end{cases}$$

其中

$$\boldsymbol{u}_j = (u_{0,j}, u_{1,j}, \cdots, u_{N,j})^{\mathrm{T}}$$

$$\boldsymbol{g}_{j+1} = (-h\mu_1((j+1)\tau), 0, \cdots, 0, h\mu_2((j+1)\tau))^{\mathrm{T}}$$

$$P=\begin{bmatrix} c_{j+1} & -1 & & & \\ -ar & p & -ar & & \\ & \ddots & \ddots & \ddots & \\ & & -ar & p & -ar \\ & & & -1 & d_{j+1} \end{bmatrix}$$

$$Q=\begin{bmatrix} 0 & 0 & & & \\ ar & q & ar & & \\ & \ddots & \ddots & \ddots & \\ & & ar & q & ar \\ & & & 0 & 0 \end{bmatrix}$$

$$c_{j+1}=1+h\lambda_1((j+1)\tau),\quad p=2(1+ar)$$
$$d_{j+1}=1+h\lambda_2((j+1)\tau),\quad q=2(1-ar)$$

4. Richardson 格式

利用公式(8.23)和(8.24)，可得到式(8.17)在内节点(k,j)处的又一差分近似

$$u_{k,j+1}=u_{k,j-1}+2ar(u_{k-1,j}-2u_{k,j}+u_{k+1,j})\tag{8.45}$$

它的截断误差是$O(\tau^2+h^2)$，因而具有二阶精度。这种差分方程属于三层显式格式，称为 Richardson(李查逊)格式，它用到的节点如图 8-10 所示。

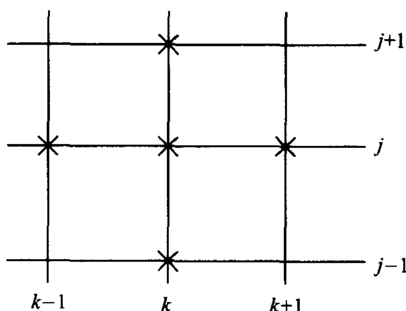


图 8-10　Richardson 格式的节点

使用差分方程(8.45)并配上离散化了的定解条件可求解初值问题(8.15)、(8.16)以及初边值问题(8.17)、(8.18)、(8.19)或(8.20)。由于式(8.45)是三层格式，一开始就要由$j=0$和$j=1$两个时间层上的u值计算$j=2$时间层上的u。因此，需要用其他方法事先给出$j=1$时间层各内节点处的u值。

5. 三层隐式格式

利用 Taylor 级数容易证明下面公式成立

$$\left(\frac{\partial u}{\partial t}\right)_{(k,j+1)}=\frac{3}{2}\left(\frac{\partial u}{\partial t}\right)_{(k,j+\frac{1}{2})}-\frac{1}{2}\left(\frac{\partial u}{\partial t}\right)_{(k,j-\frac{1}{2})}+O(\tau^2)=$$
$$\frac{3}{2}\,\frac{u(k,j+1)-u(k,j)}{\tau}-\frac{1}{2}\,\frac{u(k,j)-u(k,j-1)}{\tau}+O(\tau^2)$$

利用上面的公式以及公式(8.24)，可得到在内节点$(k,j+1)$处逼近偏微分方程(8.17)的又一差分方程

$$\frac{3}{2}\,\frac{u_{k,j+1}-u_{k,j}}{\tau}-\frac{1}{2}\,\frac{u_{k,j}-u_{k,j-1}}{\tau}=a\,\frac{u_{k-1,j+1}-2u_{k,j+1}+u_{k+1,j+1}}{h^2}$$

把它整理为

$$-2aru_{k-1,j+1}+(3+4ar)u_{k,j-1}-2aru_{k+1,j+1}=4u_{k,j}-u_{k,j-1}\tag{8.46}$$

差分方程(8.46)具有二阶精度，属于三层隐式格式，它用到的节点如图 8-11 所示。

如果使用差分方程(8.46)求解第一类初边值问题(8.17)、(8.18)、(8.19)，则计算格式是

$$\begin{cases} \boldsymbol{C}\boldsymbol{u}_{j+1} = 4\boldsymbol{u}_j - \boldsymbol{u}_{j-1} + 2\boldsymbol{f}_{j+1} \quad (j = 1,2,\cdots,m-1) \\ \boldsymbol{u}_0 = \boldsymbol{\phi} \\ \boldsymbol{u}_1 \text{ 另外计算} \end{cases}$$

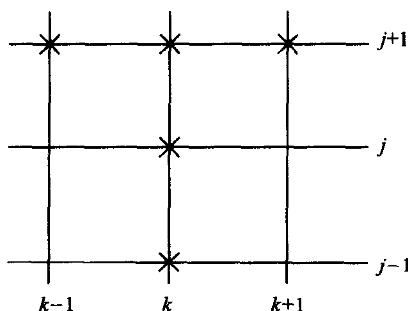


图 8-11 三层隐式格式的节点

其中矩阵 $\boldsymbol{C}$ 为

$$\boldsymbol{C} = \begin{bmatrix} 3+4ar & -2ar & & & \\ -2ar & 3+4ar & -2ar & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -2ar \\ & & & -2ar & 3+4ar \end{bmatrix}$$

$$\boldsymbol{u}_j = (u_{1,j}, u_{2,j}, \cdots, u_{N-1,j})^{\mathrm{T}}$$

$$\boldsymbol{f}_{j+1} = (ar\mu_1((j+1)\tau), 0, \cdots, 0, ar\mu_2((j+1)\tau))^{\mathrm{T}}$$

$$\boldsymbol{\phi} = (\varphi(h), \varphi(2h), \cdots, \varphi((N-1)h))^{\mathrm{T}}$$

8.2.2 差分方程的稳定性

用差分解法求解初值问题和初边值问题，每步计算总要产生舍入误差。由于求解过程是逐层计算的，属于递推算法，因此误差还会传播。若误差在传播过程中越来越大，得不到控制，则最终会把差分方程的真解给淹没了，这就是数值不稳定性。对于一个不具备数值稳定性的差分方程尽管会有一些优点也是没有实用价值的。

在研究差分方程的数值稳定性时，总假定只在初始层的节点上 $u_{k,0}$ 产生误差 $\varepsilon_{k,0}$，实际值为 $\tilde{u}_{k,0} = u_{k,0} - \varepsilon_{k,0}$，而边界的计算是精确的，又假定各层的计算都是精确进行的。在以上的假定下，考察初始误差 $\varepsilon_{k,0}$ 的传播是否得到控制。

记

$$(Lu)_{(k,j)} = \left(\frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial x^2}\right)_{(k,j)}$$

$(Lu)_{(k,j)}$ 的差分近似记为 $L_{h,\tau}u(k,j)$，例如，$(Lu)_{(k,j)}$ 的一种差分近似为

$$L_{h,\tau}u(k,j) = \frac{u(k,j+1) - u(k,j)}{\tau} - a\frac{u(k+1,j) - 2u(k,j) + u(k-1,j)}{h^2}$$

于是，逼近偏微分方程(8.14)的差分方程统一表示为

$$L_{h,\tau}u_{k,j} = 0 \tag{8.47}$$

而

$$R_{k,j} = (Lu)_{(k,j)} - L_{h,\tau}u(k,j)$$

就是差分方程(8.47)对于偏微分方程(8.14)的截断误差。

根据假定，$u_{k,0}$ 没有误差时，差分方程(8.47)的解 $u_{k,j}$ 应满足

$$\begin{cases} L_{h,\tau}u_{k,j} = 0 \\ u_{k,0} \text{ 已知} \end{cases} \tag{8.48}$$

$u_{k,0}$ 有误差时，差分方程的解 $\tilde{u}_{k,j}$ 应满足

$$\begin{cases} L_{h,\tau}\tilde{u}_{k,j} = 0 \\ \tilde{u}_{k,0} = u_{k,0} - \varepsilon_{k,0} \end{cases} \tag{8.49}$$

对于初值问题，记

$$\boldsymbol{\varepsilon}_j = (\cdots, \varepsilon_{-2,j}, \varepsilon_{-1,j}, \varepsilon_{0,j}, \varepsilon_{1,j}, \varepsilon_{2,j}, \cdots)^{\mathrm{T}}$$

其中 $\varepsilon_{k,j} = u_{k,j} - \tilde{u}_{k,j}$。

对于初边值问题，则

$$\boldsymbol{\varepsilon}_j = (\varepsilon_{1,j}, \varepsilon_{2,j}, \cdots, \varepsilon_{N-1,j})^{\mathrm{T}}$$

定义　若对于差分方程(8.48)，存在一个与 h, τ 无关的常数 K，使得当 $\tau \leqslant \tau_0, j\tau \leqslant T$ 时，$\boldsymbol{\varepsilon}_j$ 的某种范数满足

$$\| \boldsymbol{\varepsilon}_j \| \leqslant K \| \boldsymbol{\varepsilon}_0 \|, \quad j \geqslant 1$$

则称差分方程(8.48)是稳定的。

研究差分方程的稳定性的方法很多，这里仅介绍一种常用的方法，就是 Fourier(傅里叶)方法，简称傅氏方法。它适用于常系数线性差分方程。

由于式(8.14)是常系数线性偏微分方程，因而 L 和 $L_{h,\tau}$ 都是常系数线性算子。由式(8.48)与(8.49)相减，得到

$$\begin{cases} L_{h,\tau}\varepsilon_{k,j} = 0 \\ \varepsilon_{k,0} \text{ 已知} \end{cases} \tag{8.50}$$

称式(8.50)为差分方程(8.48)的误差方程。根据假定，初始层存在一个误差函数 $\varepsilon(x,0)$。对于初值问题，$\varepsilon(x,0)$ 可表示为 Fourier 积分的形式

$$\varepsilon(x,0) = \int_{-\infty}^{\infty} D_n \mathrm{e}^{\mathrm{i}nx} \mathrm{d}n$$

对于初边值问题，$\varepsilon(x,0)(0 \leqslant x \leqslant l)$ 可表示为 Fourier 级数的形式

$$\varepsilon(x,0) = \sum_{p=-\infty}^{\infty} c_p \mathrm{e}^{\mathrm{i}p\omega x}$$

其中 $\omega = \frac{\pi}{l}$。任意取出一个简谐波 $c_p \mathrm{e}^{\mathrm{i}p\omega x}$(或 $D_n \mathrm{e}^{\mathrm{i}nx}$)，并且仍记为

$$\varepsilon(x,0) = c_p \mathrm{e}^{\mathrm{i}p\omega x} (\text{或 } D_n \mathrm{e}^{\mathrm{i}nx})$$

由于只考虑初始误差的传播是否得到控制，并不计算误差的实际大小，因此可视 $c_p = 1$(或 $D_n = 1$)，并记 $n = p\omega$。于是，在稳定性的讨论中，视初始误差为振幅等于 1、频率等于 n 的简谐波

$$\varepsilon(x,0) = \mathrm{e}^{\mathrm{i}nx}$$

因而在初始层节点处就有

$$\varepsilon_{k,0} = \mathrm{e}^{\mathrm{i}nkh}$$

由于 $L_{h,\tau}$ 是线性算子，故 $\varepsilon_{k,0}$ 在传播过程中频率不会改变，故可设

$$\varepsilon_{k,j} = G^j \mathrm{e}^{\mathrm{i}nkh} \tag{8.51}$$

把式(8.51)代入误差方程(8.50)，可得到一个关于 G 的代数方程，称为差分方程(8.48)的特征方程。由于式(8.48)是常系数线性差分方程，因而它的特征方程的根 G 只依赖于 n, h, τ，而与 k, j 无关。$|G|$ 称为增长因子。由式(8.51)可知

$$\boldsymbol{\varepsilon}_j = G^j \boldsymbol{\varepsilon}_0$$

$$\| \boldsymbol{\varepsilon}_j \| = |G|^j \| \boldsymbol{\varepsilon}_0 \|$$

若 $|G| \leqslant 1$，则 $\| \boldsymbol{\varepsilon}_j \| \leqslant \| \boldsymbol{\varepsilon}_0 \| (j \geqslant 1)$；若 $|G| > 1$，则不存在 K，使 $\| \boldsymbol{\varepsilon}_j \| \leqslant K \| \boldsymbol{\varepsilon}_0 \| (j \geqslant 1)$ 成立。此外，由于初始误差是不同频率的谐波迭加，并且由于计算中舍入误差的随机性，应该认为任

何频率 n 的简谐波都是可能出现的。因此，差分方程(8.48)稳定的充分必要条件是：对任何实数 n，$|G|\leqslant 1$ 成立。

例 2 讨论差分方程(8.26)的稳定性。

解 差分方程(8.26)的误差方程为

$$\varepsilon_{k,j+1} = ar\varepsilon_{k-1,j} + (1-2ar)\varepsilon_{k,j} + ar\varepsilon_{k+1,j}$$

把式(8.51)代入上式，得

$$G^{j+1}e^{inkh} = arG^{j}e^{in(k-1)h} + (1-2ar)G^{j}e^{inkh} + arG^{j}e^{in(k+1)h}$$

消去公因子 $G^{j}e^{inkh}$，得到式(8.26)的特征方程

$$G = are^{-inh} + (1-2ar) + are^{inh} \tag{8.52}$$

利用公式

$$\cos nh = \frac{1}{2}(e^{-inh} + e^{inh})$$

把特征方程(8.52)的根整理成

$$G = 1 - 2ar(1-\cos nh) = 1 - 4ar\sin^2\frac{nh}{2}$$

要使 $|G|\leqslant 1$ 对任何实数 n 成立，必须且只须对任何实数 n 有

$$2ar\sin^2\frac{nh}{2} \leqslant 1$$

因而要求 $ar\leqslant\frac{1}{2}$。所以，差分方程(8.26)稳定的条件是 $r\leqslant\frac{1}{2a}$，即

$$\frac{\tau}{h^2} \leqslant \frac{1}{2a}$$

故称差分方程(8.26)是条件稳定的。

例 3 讨论差分方程(8.39)的稳定性。

解 把式(8.51)代入差分方程(8.39)的误差方程，并消去公因子 $G^{j}e^{inkh}$，得到式(8.39)的特征方程

$$-are^{-ink} + (1+2ar) - are^{ink} = G^{-1}$$

解出 G，得

$$G = \frac{1}{1 + 4ar\sin^2\frac{nh}{2}}$$

显然，对任何 τ,h 和实数 n，都有 $|G|\leqslant 1$。故称差分方程(8.39)是无条件稳定的。

例 4 讨论差分方程(8.45)的稳定性。

解 把式(8.51)代入式(8.45)的误差方程，经化简，得到式(8.45)的特征方程

$$G^2 + \left(8ar\sin^2\frac{nh}{2}\right)G - 1 = 0$$

解得

$$G_{1,2} = -4ar\sin^2\frac{nh}{2} \pm \sqrt{\left(4ar\sin^2\frac{nh}{2}\right)^2 + 1}$$

考察 G_2，

$$G_2 = -4ar\sin^2\frac{nh}{2} - \sqrt{\left(4ar\sin^2\frac{nh}{2}\right)^2 + 1}$$

不论 h,τ 如何选取，总存在 n，使 $|G_2|>1$，因此，Richardson 差分格式(8.45)恒不稳定，故又称为无条件不稳定。

8.3　双曲型方程的特征-差分解法

8.3.1　一阶双曲型方程

双曲型方程定解问题最典型的是一阶常系数双曲型方程初值问题

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0, \quad 0 < t < T, -\infty < x < \infty & (8.53) \\ u(x,0) = \varphi(x), \quad -\infty < x < \infty & (8.54) \end{cases}$$

其中 a 是非零常数。

容易验证，初值问题(8.53)、(8.54)的解为

$$u(x,t) = \varphi(x - at), \quad 0 \leqslant t \leqslant T, -\infty < x < \infty \tag{8.55}$$

由此看出，在 Oxt 坐标平面上，沿直线

$$x - at = c(\text{常数}) \tag{8.56}$$

u 的值保持不变，即 $u(x,t)=\varphi(c)$。直线(8.56)称为一阶双曲型方程(8.53)的特征线，它与 x 轴相交于坐标为 c 的点。当 $a>0$ 时，特征线往右上方倾斜；当 $a<0$ 时，特征线往左上方倾斜，见图 8-12。

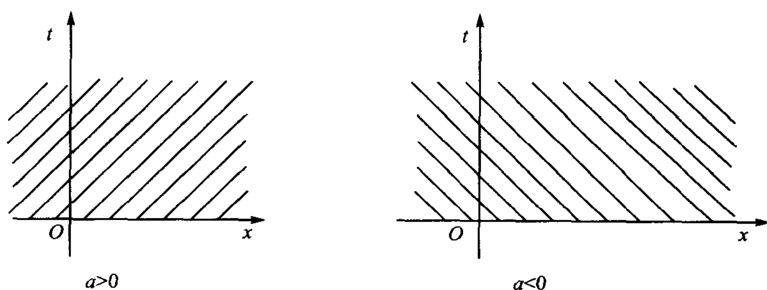


图 8-12　一阶双曲型方程的特征线

一阶双曲型方程的另一类定解问题是初边值问题，它的求解区域为 $\{0\leqslant t\leqslant T, 0\leqslant x\leqslant l\}$，其初始条件是

$$u(x,0) = \varphi(x), \quad 0 \leqslant x \leqslant l \tag{8.57}$$

由于 $u(x,t)$ 的值沿着特征线保持不变，因此，当 $a>0$ 时，在右边界 $x=l$ 上不能给出边界条件，只能在左边界 $x=0$ 上给出边界条件

$$u(0,t) = \mu(t), \quad 0 \leqslant t \leqslant T \tag{8.58}$$

并且要满足

$$\mu(0) = \varphi(0)$$

$$\mu'(0) + a\varphi'(0) = 0$$

当 $a<0$ 时，则在左边界 $x=0$ 上不能给出边界条件，只能在右边界 $x=l$ 上给出边界条件

$$u(l,t)=\beta(t),\quad 0\leqslant t\leqslant T \tag{8.59}$$

并且要满足

$$\beta(0)=\varphi(l)$$
$$\beta'(0)+a\varphi'(l)=0$$

容易验证，初边值问题

$$\begin{cases}\dfrac{\partial u}{\partial t}+a\dfrac{\partial u}{\partial x}=0\ (a>0), & 0<t<T,0<x<l \qquad (8.60)\\ u(x,0)=\varphi(x), & 0\leqslant x\leqslant l \qquad (8.57)\\ u(0,t)=\mu(t), & 0\leqslant t\leqslant T \qquad (8.58)\end{cases}$$

的解为

$$u(x,t)=\begin{cases}\varphi(x-at), & 0\leqslant t<\dfrac{x}{a},0\leqslant x\leqslant l\\ \mu\left(t-\dfrac{x}{a}\right), & \dfrac{x}{a}\leqslant t\leqslant T,0\leqslant x\leqslant l\end{cases} \tag{8.61}$$

而初边值问题

$$\begin{cases}\dfrac{\partial u}{\partial t}+a\dfrac{\partial u}{\partial x}=0(a<0), & 0<t<T,0<x<l \qquad (8.62)\\ u(x,0)=\varphi(x), & 0\leqslant x\leqslant l \qquad (8.57)\\ u(l,t)=\beta(t), & 0\leqslant t\leqslant T \qquad (8.59)\end{cases}$$

的解为

$$u(x,t)=\begin{cases}\varphi(x-at), & 0\leqslant t<\dfrac{x-l}{a},0\leqslant x\leqslant l\\ \beta\left(t-\dfrac{x-l}{a}\right), & \dfrac{x-l}{a}\leqslant t\leqslant T,0\leqslant x\leqslant l\end{cases} \tag{8.63}$$

现在考虑右端不恒为零的一阶双曲型方程

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=f(x,t,u) \tag{8.64}$$

它的特征线仍为式(8.56)。由于当 $x=c+at$ 时，有

$$u(x,t)=u(c+at,t),\quad \frac{\mathrm{d}u}{\mathrm{d}t}=\frac{\partial u}{\partial x}a+\frac{\partial u}{\partial t}$$

所以在特征线(8.56)上，方程(8.64)成为

$$\frac{\mathrm{d}u}{\mathrm{d}t}=f(x,t,u) \tag{8.65}$$

因此，初值问题(8.64)、(8.54)在通过 x 轴上任一点 $Q(x_Q,0)$ 的特征线 $x=x_Q+at$ 上就成为常微分方程初值问题

$$\begin{cases}\dfrac{\mathrm{d}u}{\mathrm{d}t}=f(x,t,u), & t>0\\ u(x_Q,0)=\varphi(x_Q)\end{cases} \tag{8.66}$$

这就可以用第 7 章所述的数值解法求出 $u(x,t)$。

设 $a>0$，见图 8-13，在 x 轴上任取一点 $Q(x_Q,0)$，过点 Q 作特征线 $x=x_Q+at$ 与直线

$t=\tau$交于点 $P_1(x_{P_1},\tau)$，其中 $x_{P_1}=x_Q+a\tau$。沿线段 QP_1 微分方程(8.65)成立。用 Euler 法求解初值问题(8.66)，可得第一层 $t=\tau$ 上的值

$$u_{P_1}=u_Q+\tau f(x_Q,0,u_Q) \tag{8.67}$$

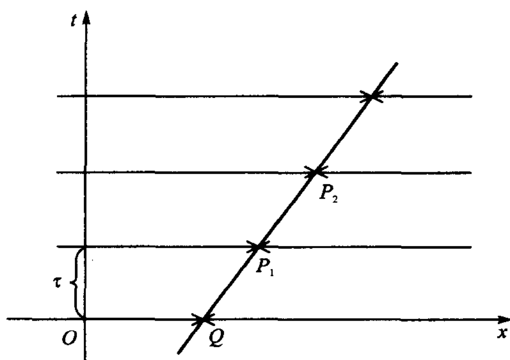


图 8-13　特征线法的节点

其中 $u_Q=\varphi(x_Q)$。当得出解在第一层的数值 u_{P_1} 后继续使用 Euler 法，按 t 的步长 τ 逐层求出 u_{P_j}，

$$u_{P_{j+1}}=u_{P_j}+\tau f(x_{P_j},j\tau,u_{P_j})\quad(j=1,2,\cdots,m-1) \tag{8.68}$$

这就是特征线法。由于用 Euler 法，故它的截断误差为 $O(\tau)$。

特征线法(8.67)、(8.68)计算过程比较简单，但由于特征线分布一般是不规则的，因而不便于编制程序。在 8.2 节介绍的差分解法有网格规则化的特性，而特征线法能反映双曲型方程的物理特性，为了同时利用两者的优点，可把二者结合起来，形成所谓特征-差分方法，其作法如下。

先把求解区域网格化，t 方向步长为 τ，x 方向步长为 h，见图8-5和图 8-6。

设 $a>0$，由点 $P(k,j+1)$引特征线与第 j 层网格线相交于点 Q，见图 8-14，利用点 $B(k-1,j)$和点 $C(k,j)$作线性插值求 u_Q(这时所用的步长 h 应保证点 B 在点 Q 的左边)，可得

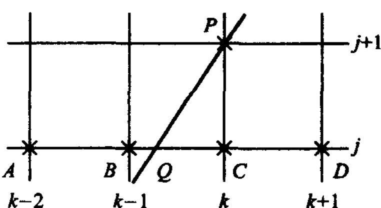


图 8-14　特征-差分方法的节点

$$u_Q=u_B+\frac{u_C-u_B}{x_k-x_{k-1}}(x_Q-x_{k-1})$$

由于 $x_Q=x_k-a\tau$，故由上式得到

$$u_Q=u_{k-1,j}+\frac{u_{k,j}-u_{k-1,j}}{h}(h-a\tau)=(1-ar)u_{k,j}+ar\,u_{k-1,j} \tag{8.69}$$

其中 $r=\dfrac{\tau}{h}$。用同样的方法得

$$f(x_Q,t_j,u_Q)=(1-ar)f_{k,j}+arf_{k-1,j} \tag{8.70}$$

其中 $f_{k,j}=f(x_k,t_j,u_{k,j})$。把式(8.69)和式(8.70)代人 Euler 法计算公式

$$u_P=u_Q+\tau f(x_Q,t_j,u_Q) \tag{8.71}$$

得到逼近微分方程(8.64)的差分方程

$$u_{k,j+1}=(1-ar)u_{k,j}+aru_{k-1,j}+\tau[(1-ar)f_{k,j}+arf_{k-1,j}] \tag{8.72}$$

为保证点 B 在点 Q 的左边，要求 $|QC|\leqslant h$，而 $|QC|=a\tau$，故要求

$$\frac{a\tau}{h} \leqslant 1$$

当函数 f 不含 u 时，此条件是差分方程(8.72)稳定的条件，证明从略。

若 $a<0$，则用类似的方法可得逼近微分方程(8.64)的差分方程

$$u_{k,j+1} = (1+ar)u_{k,j} - aru_{k+1,j} + \tau[(1+ar)f_{k,j} - arf_{k+1,j}] \tag{8.73}$$

当函数 f 不含 u 时，它稳定的条件为

$$\frac{|a|\tau}{h} \leqslant 1$$

如果用点 B 和点 D 作线性插值求 u_Q 和 $f(x_Q,t_j,u_Q)$，则可得

$$u_Q = u_B + \frac{u_D - u_B}{x_{k+1} - x_{k-1}}(x_Q - x_{k-1}) = \frac{1}{2}(1+ar)u_{k-1,j} + \frac{1}{2}(1-ar)u_{k-1,j}$$

$$f(x_Q,t_j,u_Q) = \frac{1}{2}(1+ar)f_{k-1,j} + \frac{1}{2}(1-ar)f_{k+1,j}$$

代入 Euler 法计算公式(8.71)得

$$u_{k,j+1} = \frac{1}{2}(1+ar)u_{k-1,j} + \frac{1}{2}(1-ar)u_{k+1,j} + \frac{\tau}{2}[(1+ar)f_{k-1,j} + (1-ar)f_{k+1,j}] \tag{8.74}$$

用差分方程(8.74)逼近微分方程(8.64)，无论 $a>0$ 和 $a<0$ 都适用。当函数 f 不含 u 时，它稳定的条件为 $\frac{|a|\tau}{h}\leqslant 1$。

用差分方程(8.72)求解初值问题(8.64)($a>0$)、(8.54)的计算格式为

$$\begin{cases} u_{k,j+1} = (1-ar)u_{k,j} + aru_{k-1,j} + \tau[(1-ar)f_{k,j} + arf_{k-1,j}] \\ \qquad (k=0,\pm 1,\pm 2,\cdots;j=0,1,\cdots,m-1) \\ u_{k,0} = \varphi(kh) \quad (k=0,\pm 1,\pm 2,\cdots) \end{cases}$$

而求解初边值问题(8.64)($a>0$)、(8.57)、(8.58)的计算格式为

$$\begin{cases} u_{k,j+1} = (1-ar)u_{k,j} + aru_{k-1,j} + \tau[(1-ar)f_{k,j} + arf_{k-1,j}] \\ \qquad (k=1,2,\cdots,N;j=0,1,\cdots,m-1) \\ u_{k,0} = \varphi(kh) \quad (k=1,2,\cdots,N) \\ u_{0,j} = \mu(j\tau) \quad (j=0,1,\cdots,m) \end{cases}$$

当用差分方程(8.74)求解初边值问题时，对于不给边界条件的那条边界上节点的 u 值，需要用另外的方法求出。例如求解初边值问题(8.64)($a>0$)、(8.57)、(8.58)，当第 j 层上的 $u_{1,j},u_{2,j},\cdots,u_{N-1,j}$ 用格式(8.74)计算完后，$u_{N,j}$ 无法用格式(8.74)计算。这时，可用格式(8.72)计算 $u_{N,j}$，得

$$u_{N,j} = (1-ar)u_{N,j-1} + aru_{N-1,j-1} + \tau[(1-ar)f_{N,j-1} + arf_{N-1,j-1}]$$

或以($N-2,j$)和($N-1,j$)为插值节点进行线性外推计算 $u_{N,j}$，得

$$u_{N,j} = 2u_{N-1,j} - u_{N-2,j}$$

于是，用格式(8.74)求解初边值问题(8.64)($a>0$)、(8.57)、(8.58)，并用线性外推法处理右边界问题，得计算格式如下：

$$\begin{cases}\begin{cases}\text{差分方程(8.74)} & (k=1,2,\cdots,N-1)\\ u_{N,j+1}=2u_{N-1,j+1}-u_{N-2,j+1} & (j=0,1,\cdots,m-1)\end{cases}\\ u_{k,0}=\varphi(kh) \quad (k=1,2,\cdots,N)\\ u_{0,j}=\mu(j\tau) \quad (j=0,1,\cdots,m)\end{cases}$$

8.3.2 一阶双曲型方程组

设 $\boldsymbol{u}=(u^{(1)},u^{(2)},\cdots,u^{(p)})^{\mathrm{T}}$ 是 x 和 t 的向量值函数，$\boldsymbol{A}$ 是 $p\times p$ 的常数矩阵，它的所有特征值 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 都是实数，并且存在非奇异矩阵 $\boldsymbol{Q}$，使得

$$\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q}=\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p) \tag{8.75}$$

则称偏微分方程组

$$\frac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{A}\frac{\partial \boldsymbol{u}}{\partial x}=\boldsymbol{f}(x,t,\boldsymbol{u}) \tag{8.76}$$

为一阶双曲型方程组，其中 $\boldsymbol{f}$ 是 p 维向量

$$\boldsymbol{f}(x,t,\boldsymbol{u})=(f_1(x,t,\boldsymbol{u}),\cdots,f_p(x,t,\boldsymbol{u}))^{\mathrm{T}}$$

如果 $\boldsymbol{A}$ 的特征值互异，则称式(8.76)为严格双曲型方程组。

一阶双曲型方程组的初值问题为

$$\begin{cases}\dfrac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{A}\dfrac{\partial \boldsymbol{u}}{\partial x}=\boldsymbol{f}(x,t,\boldsymbol{u}), \quad 0<t<T,-\infty<x<\infty & (8.77)\\ \boldsymbol{u}(x,0)=\boldsymbol{\Phi}(x), \quad -\infty<x<\infty & (8.78)\end{cases}$$

其中 $\boldsymbol{\Phi}(x)=(\varphi_1(x),\varphi_2(x),\cdots,\varphi_p(x))^{\mathrm{T}}$ 是 x 的向量值函数。

可使用差分格式(8.72)、(8.73)或(8.74)求解初值问题(8.77)、(8.78)，但使用格式(8.74)更方便一些。把式(8.74)用于方程组(8.77)的情形，得以下的差分方程

$$\begin{aligned}\boldsymbol{u}_{k,j+1}=&\frac{1}{2}(\boldsymbol{I}+r\boldsymbol{A})\boldsymbol{u}_{k-1,j}+\frac{1}{2}(\boldsymbol{I}-r\boldsymbol{A})\boldsymbol{u}_{k+1,j}+\\ &\frac{\tau}{2}[(\boldsymbol{I}+r\boldsymbol{A})\boldsymbol{f}_{k-1,j}+(\boldsymbol{I}-r\boldsymbol{A})\boldsymbol{f}_{k+1,j}]\end{aligned} \tag{8.79}$$

其中 $\boldsymbol{I}$ 是 $p\times p$ 单位矩阵，向量 $\boldsymbol{f}_{k,j}=(f_1(x_k,t_j,\boldsymbol{u}_{k,j}),\cdots,f_p(x_k,t_j,\boldsymbol{u}_{k,j}))^{\mathrm{T}}$。利用格式(8.79)求解初值问题(8.77)、(8.78)的计算格式为

$$\begin{cases}\text{差分方程(8.79)} \quad (k=0,\pm1,\pm2,\cdots;j=0,1,\cdots,m-1)\\ \boldsymbol{u}_{k,0}=\boldsymbol{\Phi}(kh) \quad (k=0,\pm1,\pm2,\cdots)\end{cases}$$

8.3.3 二阶双曲型方程

最典型的二阶双曲型方程是波动方程

$$Lu\equiv\frac{\partial^2 u}{\partial t^2}-a^2\frac{\partial^2 u}{\partial x^2}=f(x,t) \tag{8.80}$$

其中 $a>0$。现仅考虑它的初值问题

$$\begin{cases}Lu=f(x,t), \quad 0<t<T,-\infty<x<\infty & (8.80)\\ u(x,0)=\varphi(x), \quad \dfrac{\partial u}{\partial t}\Big|_{t=0}=\psi(x), \quad -\infty<x<\infty & (8.81)\end{cases}$$

式(8.81)称为初始条件。

在式(8.80)中令

$$\omega_1 = \frac{\partial u}{\partial t}, \quad \omega_2 = a\frac{\partial u}{\partial x} \tag{8.82}$$

并利用 $\frac{\partial^2 u}{\partial x \partial t} = \frac{\partial^2 u}{\partial t \partial x}$，可得一阶双曲型方程组

$$\begin{cases} \dfrac{\partial \omega_1}{\partial t} - a\dfrac{\partial \omega_2}{\partial x} = f(x,t) \\ \dfrac{\partial \omega_2}{\partial t} - a\dfrac{\partial \omega_1}{\partial x} = 0 \end{cases}$$

令 $\boldsymbol{\omega}=(\omega_1,\omega_2)^{\mathrm{T}}$，则上述方程组可表示为

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \boldsymbol{A}\frac{\partial \boldsymbol{\omega}}{\partial x} = \boldsymbol{f} \tag{8.83}$$

其中 $\boldsymbol{f}=(f(x,t),0)^{\mathrm{T}}$，矩阵

$$\boldsymbol{A} = \begin{bmatrix} 0 & -a \\ -a & 0 \end{bmatrix}$$

易知，$\boldsymbol{A}$ 的特征值为 $\lambda_1=a,\lambda_2=-a$。取

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} a & 0 \\ 0 & -a \end{bmatrix}$$

则有 $\boldsymbol{A}=\boldsymbol{Q\Lambda Q}^{-1}$，故式(8.83)是严格双曲型方程组。求解二阶双曲型方程初值问题(8.80)、(8.81)可转化为求解与之等价的一阶双曲型方程组初值问题

$$\begin{cases} \dfrac{\partial \boldsymbol{\omega}}{\partial t} + \boldsymbol{A}\dfrac{\partial \boldsymbol{\omega}}{\partial x} = \boldsymbol{f}, \quad 0<t<T, -\infty<x<\infty & (8.83) \\ \boldsymbol{\omega}(x,0) = (\psi(x), a\varphi'(x))^{\mathrm{T}}, \quad -\infty<x<\infty & (8.84) \end{cases}$$

如果使用差分格式(8.79)，则计算格式为

$$\begin{cases} \boldsymbol{\omega}_{k,j+1} = \dfrac{1}{2}(\boldsymbol{I}+r\boldsymbol{A})\boldsymbol{\omega}_{k-1,j} + \dfrac{1}{2}(\boldsymbol{I}-r\boldsymbol{A})\boldsymbol{\omega}_{k+1,j} + \\ \qquad \dfrac{\tau}{2}[(\boldsymbol{I}+r\boldsymbol{A})\boldsymbol{f}_{k-1,j} + (\boldsymbol{I}-r\boldsymbol{A})\boldsymbol{f}_{k+1,j}] \\ \qquad (k=0,\pm1,\pm2,\cdots;j=0,1,\cdots,m-1) \\ \boldsymbol{\omega}_{k,0} = (\psi(kh), a\varphi'(kh))^{\mathrm{T}} \quad (k=0,\pm1,\pm2,\cdots) \end{cases}$$

求出 $\boldsymbol{\omega}=(\omega_1,\omega_2)^{\mathrm{T}}$ 之后，再由式(8.82)，得

$$u(x,t) = u(x,0) + \int_0^t \omega_1(x,t)\mathrm{d}t$$

或

$$u(k,j) = \varphi(kh) + \int_0^{j\tau} \omega_1(kh,t)\mathrm{d}t$$

用复化梯形公式计算上式右端的积分，可得

$$u_{k,j} = \varphi(kh) + \frac{\tau}{2}\left(\omega_{1,k,0} + \omega_{1,k,j} + 2\sum_{i=1}^{j-1}\omega_{1,k,i}\right) \tag{8.85}$$

$$(k=0,\pm1,\pm2,\cdots;j=1,2,\cdots,m)$$

其中 $\omega_{1,k,i}$是向量 $\boldsymbol{\omega}_{k,i}$的第 1 个分量。式(8.85)就是初值问题(8.80)、(8.81)的数值解。

习　　题

1. 证明下列公式成立：

(1) $\left(\dfrac{\partial u}{\partial x}\right)_{(k,j)}=\dfrac{1}{2h}[u(k+1,j)-u(k-1,j)]+O(h^2)$；

(2) $\left(\dfrac{\partial^2 u}{\partial x^2}\right)_{(k,j)}=\dfrac{1}{h^2}[u(k+1,j)-2u(k,j)+u(k-1,j)]+O(h^2)$。

2. 试用差分格式(8.8)求解 Poisson 方程的边值问题

$$\begin{cases}\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=16, & (x,y)\in\Omega\\ u\,|_\Gamma=0, & (x,y)\in\Gamma\end{cases}$$

其中 $\Omega=\{(x,y)\,\big|\,|x|<1,|y|<1\}$，分别取步长 $h=1$ 和 $h=0.5$ 求解。

3. 设区域 $\Omega+\Gamma$ 如图 8－15 所示。

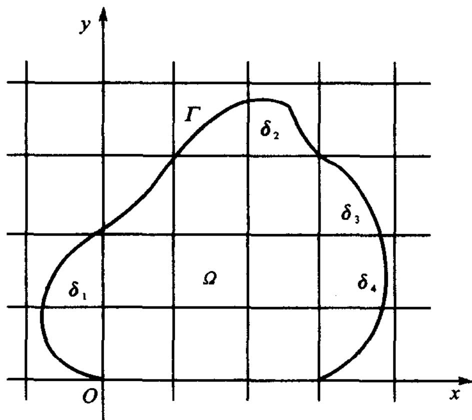


图 8－15　习题 3 附图

其中网格为正方形，步长为 h，试写出求解 Poisson 方程边值问题

$$\begin{cases}\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=f(x,y), & (x,y)\in\Omega\\ u\,|_\Gamma=\varphi(x,y), & (x,y)\in\Gamma\end{cases}$$

的五点差分格式，并写出其矩阵形式。

4. 试用差分格式(8.8)求解 Laplace 方程边值问题

$$\begin{cases}\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=0, \quad 0<x<4, \quad 0<y<3\\ u(0,y)=y(3-y), \quad u(4,y)=0, \quad 0\leqslant y\leqslant 3\\ u(x,0)=\sin\dfrac{\pi x}{4}, \quad u(x,3)=0, \quad 0\leqslant x\leqslant 4\end{cases}$$

取步长 $h=1$。

5. 求逼近扩散方程 $\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}$ 的差分格式

$$(1+\theta)\frac{u_{k,j+1}-u_{k,j}}{\tau}-\theta\frac{u_{k,j}-u_{k,j-1}}{\tau}=\frac{1}{h^2}(u_{k+1,j}-2u_{k,j}+u_{k-1,j})$$

的精度，并调整 θ 使其精度达到二阶。

6. 试用古典显式格式(8.26)求解下列定解问题

(1) $$\begin{cases}\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}, & 0<x<1,\quad t>0\\ u(x,0)=4x(1-x), & 0\leqslant x\leqslant 1\\ u(0,t)=u(1,t)=0, & t\geqslant 0\end{cases}$$

只计算 $j=1,2$ 两层上的数值解，其中取 $r=\frac{1}{6}$，$h=0.2$；

(2) $$\begin{cases}\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}, & 0<x<1,\quad t>0\\ u(x,0)=4x(1-x), & 0\leqslant x\leqslant 1\\ \dfrac{\partial u(0,t)}{\partial x}=u(0,t) & \\ \dfrac{\partial u(1,t)}{\partial x}=-u(1,t) & t\geqslant 0\end{cases}$$

只计算 $j=1,2$ 两层上的数值解，其中取 $h=0.2$，$\tau=0.1$，边界条件采用式(8.30)的形式。

7. 用古典隐式格式(8.39)求解下列定解问题

$$\begin{cases}\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}, & 0<x<1,\quad 0<t<0.3\\ u(x,0)=\sin\pi x, & 0\leqslant x\leqslant 1\\ u(0,t)=u(1,t)=0, & 0\leqslant t\leqslant 0.3\end{cases}$$

其中取 $h=0.2$，$\tau=0.1$。[准确解为 $u(x,t)=e^{-2\pi t}\sin\pi x$]

8. 试证：若条件 $0<r\leqslant\frac{1}{2}$ 成立，并且初始函数 $u_{k,0}=\varphi(kh)(k=0,1,\cdots,N)$ 对一切 k 满足不等式 $|\varphi(kh)|<M$(常数)，则差分方程

$$u_{k,j+1}=(1-2r)u_{k,j}+r(u_{k+1,j}+u_{k-1,j})$$
$$(k=1,2,\cdots,N-1;j=0,1,\cdots,m-1)$$

在齐次边界条件下的解对一切 $k(0\leqslant k\leqslant N)$ 及一切 $j\geqslant 0$ 有如下的估计式：$|u_{k,j}|<M$。

9. 讨论逼近微分方程 $\frac{\partial u}{\partial t}=a\frac{\partial^2 u}{\partial x^2}(a>0)$ 的 Crank -Nicholson格式(8.42)的稳定性。

10. 讨论逼近微分方程 $\frac{\partial u}{\partial t}=a\frac{\partial^2 u}{\partial x^2}(a>0)$ 的三层隐式格式(8.46)的稳定性。

11. 分别使用下列两组网格节点

(1) $(k,j+1),(k-1,j),(k+1,j),(k,j-1)$；

(2) $(k-1,j+1),(k,j+1),(k,j)$。

构造逼近微分方程 $\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0$ 的差分格式，并讨论它的精度和稳定性。

12. 试用特征线法求定解问题

$$\begin{cases} \dfrac{\partial u}{\partial t} + 2\dfrac{\partial u}{\partial x} = tu(x+u) - 3, & 0 < x < 10, \quad t > 0 \\ u(x,0) = e^{-x}, & 0 \leqslant x \leqslant 10 \\ u(0,t) = \dfrac{1}{t+1}, & t \geqslant 0 \end{cases}$$

在点 $p_i(i+1, 0.5i)$ 和 $q_i(i, 1+0.5i)(i=1,2,3)$ 处的数值解。(取 t 方向步长 $\tau=0.5$)

13. 试用特征线法求定解问题

$$\begin{cases} \dfrac{\partial u}{\partial t} - 3\dfrac{\partial u}{\partial x} = 4(x+u), & 0 < x < 6, \quad t > 0 \\ u(x,0) = x^2 - 36, & 0 \leqslant x \leqslant 6 \\ u(6,t) = 2t, & t \geqslant 0 \end{cases}$$

在点 $p_1(5.4, 1.2)$, $p_2(4.8, 1.4)$ 和 $p_3(4.2, 1.6)$ 处的数值解。(取 t 方向步长 $\tau=0.2$)

14. 用特征-差分方法推出逼近微分方程

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = f(x,t,u) \quad (a < 0)$$

的差分方程(8.73)。

15. 利用差分方程(8.73)求解定解问题

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{3}{8}\dfrac{\partial u}{\partial x}, & 0 < x < 1, \quad 0 < t < 0.3 \\ u(x,0) = -0.4(x-0.5)^2 + 0.1, & 0 \leqslant x \leqslant 1 \\ u(1,t) = t(t-0.3), & 0 \leqslant t \leqslant 0.3 \end{cases}$$

取 x 方向步长 $h=0.2$, t 方向步长 $\tau=0.1$。

16. 利用差分方程(8.79)求解方程组初值问题

$$\begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{A}\dfrac{\partial \boldsymbol{u}}{\partial x} = \boldsymbol{0}, & 0 < t < 0.3, -\infty < x < \infty \\ \boldsymbol{u}(x,0) = (x(x-1), 0)^{\mathrm{T}}, & -\infty < x < \infty \end{cases}$$

取 $h=0.2$, $\tau=0.1$, 只计算 $-0.4 \leqslant x \leqslant 0.4$ 范围内的解,其中

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}^{(1)} \\ \boldsymbol{u}^{(2)} \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} 0 & -2 \\ -1 & 1 \end{bmatrix}$$

17. 如果利用差分方程(8.72)、(8.73)求解第 16 题的初值问题,试写出计算格式。

18. 利用差分方程(8.79)求解下列初值问题

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = \dfrac{\partial^2 u}{\partial x^2}, & 0 < t < 0.3, -\infty < x < \infty \\ u(x,0) = 1 + x^2, \quad \dfrac{\partial u}{\partial t}\Big|_{t=0} = 0, & -\infty < x < \infty \end{cases}$$

取 $h=0.2$, $\tau=0.1$, 只计算 $-0.4 \leqslant x \leqslant 0.4$ 范围内的解。

习题答案与提示

第1章

1. $\varepsilon(a)=0.000\ 005$， $\varepsilon(b)=0.005$， $\varepsilon(c)=0.000\ 5$；
   $\varepsilon_r(a)=0.037\%$， $\varepsilon_r(b)=0.041\%$， $\varepsilon_r(c)=0.042\%$。
2. a 有四位，b 有二位，c 没有有效数字。
3. $\varepsilon(a)=0.000\ 05$。
4. (1) $\varepsilon(\tilde{u})=0.001\ 71$， $\varepsilon_r(\tilde{u})=0.12\%$， $\tilde{u}=1.462\ 87$ 有三位有效数字；
   (2) $\varepsilon(\tilde{u})=0.003\ 869$， $\varepsilon_r(\tilde{u})=0.43\%$， $\tilde{u}=-0.895\ 92$ 有二位有效数字。
5. $u=\sqrt{x}-\sqrt{y}=\dfrac{x-y}{\sqrt{x}+\sqrt{y}}=0.000\ 278\ 9$；
   $u=\tan x-\tan y=\dfrac{\sin(x-y)}{\cos x\cdot\cos y}=0.001\ 005$。
6. 使用 $u=\dfrac{1}{99+70\sqrt{2}}$，此时 $\varepsilon(\tilde{u})=0.901\ 8\times10^{-4}$。
7. $a=40+50+60+90=240$， $u=1\ 340\times10^2+a=134\ 200$。
8. $\begin{cases} y_0=0.402\ 3 \\ y_n=\dfrac{1}{4}\left(\dfrac{1}{n}-y_{n-1}\right) \quad (n=1,2,\cdots,8)\end{cases}$
   $y_1=0.149\ 4$， $y_2=0.087\ 65$， $y_3=0.061\ 43$， $y_4=0.047\ 15$， $y_5=0.038\ 23$，
   $y_6=0.032\ 13$， $y_7=0.027\ 70$， $y_8=0.024\ 33$。
9. 用向量范数定义证明。
10. 用向量范数定义证明。当 $\boldsymbol{A}$ 奇异时，$\|\cdot\|_A$ 不满足定义中的第一个条件。
11. 先证明 $\|\boldsymbol{A}\|_2\leqslant\|\boldsymbol{A}\|_F$，据此证明 $\|\cdot\|_F$ 与向量范数 $\|\cdot\|_2$ 相容，最后证明 $\|\cdot\|_F$ 是矩阵范数，其中要使用等式 $\|\boldsymbol{A}\|_F^2=\sum\limits_{j=1}^{n}\|\boldsymbol{A}_j\|_2^2$，$\boldsymbol{A}_j$ 是 $\boldsymbol{A}$ 的第 j 列向量。
12. $\|\boldsymbol{x}\|_1=14$， $\|\boldsymbol{x}\|_2=2\sqrt{21}$， $\|\boldsymbol{x}\|_\infty=8$；
    $\|\boldsymbol{A}\|_1=9$， $\|\boldsymbol{A}\|_\infty=11$， $\|\boldsymbol{A}\|_F=\sqrt{79}$。
13. 利用矩阵范数定义中的第(4)个条件证明。
14. (1) 其右半部分要利用柯西-施瓦兹不等式证明；
    (2) 利用(1)证明。
15. 设给定的矩阵范数为 $\|\cdot\|_\alpha$，构造一个只与向量 $\boldsymbol{x}\in\mathbf{R}^n$ 有关的矩阵 $\boldsymbol{F}(\boldsymbol{x})\in\mathbf{R}^{n\times n}$，令 $\|\boldsymbol{x}\|_\alpha=\|\boldsymbol{F}(\boldsymbol{x})\|_\alpha$，然后证明 $\|\boldsymbol{x}\|_\alpha$ 是一种向量范数且与给定的矩阵范数相容。

第2章

1. 顺序 Gauss 消去法

$$[\boldsymbol{A},\boldsymbol{b}]\rightarrow\cdots\rightarrow\begin{bmatrix}0.5 & 1.1 & 3.1 & 6\\ 0 & -10.0 & -24.5 & -59.0\\ 0 & 0 & -12.2 & -24.6\end{bmatrix},\quad \begin{cases}x_3=2.02\\ x_2=0.95\\ x_1=-2.62\end{cases}$$

列主元素 Gauss 消去法

$$[\boldsymbol{A},\boldsymbol{b}]\rightarrow\cdots\rightarrow\begin{bmatrix}5 & 0.96 & 6.5 & 0.96\\ 0 & 4.12 & -2.24 & -0.364\\ 0 & 0 & 2.99 & 5.99\end{bmatrix},\quad \begin{cases}x_3=2.00\\ x_2=1.00\\ x_1=-2.60\end{cases}$$

2. 是式(2.8)中的初等下三角矩阵 $\boldsymbol{P}_1$，其中 $p_{i1}=-m_{i1}(i=2,3,\cdots,n)$。

3. 用式(2.16)的初等置换矩阵 $\boldsymbol{Q}_k$。

4.

统计项目	求 l_{ij}	求 u_{ij}	求 y_i	求 x_i
乘法次数	$\frac{1}{6}(n^3-3n^2+2n)$	$\frac{1}{6}(n^3-n)$	$\frac{1}{2}(n^2-n)$	$\frac{1}{2}(n^2-n)$
除法次数	$\frac{1}{2}(n^2-n)$	0	0	n

全部乘、除法总次数为$\frac{n^3}{3}+n^2-\frac{n}{3}$。

5. 系数矩阵 $\boldsymbol{A}$ 分解为

$$\boldsymbol{A}=\begin{bmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 1.71875 & 1.35826 & 1\end{bmatrix}\begin{bmatrix}3.2 & -1.4 & 1\\ 0 & 5.6 & 1\\ 0 & 0 & -7.07701\end{bmatrix}$$

方程组的解为$x_3=-1.024\ 38, x_2=-0.388\ 503, x_1=0.931\ 399$。

6. 设系数矩阵为 $\boldsymbol{A}$，右端向量为 $\boldsymbol{b}$，经过选主元的 Doolittle 分解后，得

$$\boldsymbol{QA}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0.1 & 1 & 0 & 0\\ -0.6 & 0 & 1 & 0\\ 0.2 & 0.4 & 0.8 & 1\end{bmatrix}\begin{bmatrix}10 & 5 & -5 & 6\\ 0 & 7.5 & 2.5 & 2.4\\ 0 & 0 & 5 & 4.6\\ 0 & 0 & 0 & -3.84\end{bmatrix},\quad \boldsymbol{Qb}=\begin{bmatrix}80\\ 12\\ 40\\ -50\end{bmatrix}$$

方程组的解为 $x_1=35.937\ 5, x_3=-15.462\ 5, x_2=-5.812\ 5, x_1=-18.387\ 5$。

7. $u_i=\sum\limits_{j=\max(1,i-r)}^{\min(i+s,n)} c_{i-j+s+1,j}y_j \quad (i=1,2,\cdots,n)$。

8. 系数矩阵 $\boldsymbol{A}$ 分解为

$$\boldsymbol{A}=\begin{bmatrix}8 & & & & \\ 2 & 8.250 & & & \\ & 1 & 7.879 & & \\ & & 2 & 8.254 & \\ & & & 1 & 7.879\end{bmatrix}\begin{bmatrix}1 & -0.125 & & & \\ & 1 & 0.1212 & & \\ & & 1 & -0.1269 & \\ & & & 1 & 0.1212\\ & & & & 1\end{bmatrix}$$

方程组的解为 $x_5=-0.639\ 9, x_4=0.820\ 0, x_3=1.200, x_2=-0.519\ 9, x_1=0.210\ 0$。

9. 增广矩阵变换为

$$\begin{bmatrix} 1.2969 & 0.8648 & 0.8642 \\ 0 & 10^{-8} & -10^{-8} \end{bmatrix}$$

回代后得 $x_2=-1, x_1=1.333\ 179\ 1$，此结果与精确解 $x_1=2, x_2=-2$ 相差甚远。系数矩阵 $\boldsymbol{A}$ 的条件数 $\text{cond}(\boldsymbol{A})_\infty=\|\boldsymbol{A}\|_\infty\|\boldsymbol{A}^{-1}\|_\infty=2.161\ 7\times1.513\times10^8=3.270\ 652\ 1\times10^8$，方程组病态严重，虽然使用列主元素 Gauss 消去法求解，且取八位十进制，也得不出好的结果。

10. 利用 $\boldsymbol{A}$ 的正交性证明。因 $\boldsymbol{A}$ 是正交矩阵，故 $\boldsymbol{A}^{-1}=\boldsymbol{A}^{\mathrm{T}}$。

11. 利用矩阵范数与向量范数的相容性证明。

12. (1) $\text{cond}(\boldsymbol{A})_\infty=\|\boldsymbol{A}\|_\infty\|\boldsymbol{A}^{-1}\|_\infty=1.99\times1.99\times10^4=39\ 601$；

 (2) $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}=\begin{bmatrix}0\\0.01\end{bmatrix}$；

 (3) $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\tilde{\boldsymbol{x}}=\begin{bmatrix}-0.995\\-0.985\end{bmatrix}$。

 分析：(2)的 $\|\boldsymbol{r}\|_\infty$ 小于(3)的 $\|\boldsymbol{r}\|_\infty$ 并不说明(2)的 $\tilde{\boldsymbol{x}}$ 的精度高于(3)的 $\tilde{\boldsymbol{x}}$ 的精度。这是因为系数矩阵 $\boldsymbol{A}$ 的病态比较严重，残向量的范数小不一定能说明近似解的精度高。

13. (1) 迭代矩阵 $\boldsymbol{G}$ 的谱半径 $\rho(\boldsymbol{G})=5>1$，故 $\{\boldsymbol{x}^{(k)}\}$ 不收敛；

 (2) 迭代矩阵 $\boldsymbol{G}$ 的行范数 $\|\boldsymbol{G}\|_\infty=0.9<1$，故 $\{\boldsymbol{x}^{(k)}\}$ 收敛。

14. 先求出迭代矩阵 $\boldsymbol{G}$，然后利用有关定理证明迭代公式收敛。$\boldsymbol{G}$ 和要回答的方程组分别是

$$\boldsymbol{G}=\begin{bmatrix}0 & 0.4 & 0.2\\0 & -0.9 & -0.45\\0 & -0.35 & 0.625\end{bmatrix},\quad \begin{cases}x_1-0.4x_2-0.2x_3=1\\0.25x_1-\ 2x_2-0.5x_3=-2\\0.25x_1+0.5x_2-0.2x_3=-3\end{cases}$$

15. 因系数矩阵是主对角线元素按行严格占优阵，故用下列的 Jacobi 迭代法求解必收敛：

$$\begin{cases}x_1^{(k+1)}=-0.4x_2^{(k)}-0.2x_3^{(k)}-2.4\\x_2^{(k+1)}=0.25x_1^{(k)}-0.5x_3^{(k)}+2.5\\x_3^{(k+1)}=-0.2x_1^{(k)}+0.5x_2^{(k)}+0.1\end{cases}\quad(k=0,1,2\cdots)$$

 $\boldsymbol{x}^*\approx(-3.090\ 900\ 373, 1.094\ 597\ 704, 1.265\ 449\ 512)^{\mathrm{T}}$。

16. (1)发散； (2)收敛； (3)收敛； (4)发散。

17. 用下列 Gauss－Seidel 迭代法求解必收敛(理由同第 15 题)：

$$\begin{cases}x_1^{(k+1)}=-0.4x_2^{(k)}-0.2x_3^{(k)}-2.4\\x_2^{(k+1)}=0.25x_1^{(k+1)}-0.5x_3^{(k)}+2.5\\x_3^{(k+1)}=-0.2x_1^{(k+1)}+0.5x_2^{(k+1)}+0.1\end{cases}\quad(k=0,1,2\cdots)$$

18. (1) $\boldsymbol{G}=\begin{bmatrix}0 & 0.5 & -0.5\\0 & -0.5 & -0.5\\0 & 0 & -0.5\end{bmatrix}$，$\rho(\boldsymbol{G})=0.5<1$，所以收敛；

 (2) $\boldsymbol{G}=\begin{bmatrix}0 & -2 & 2\\0 & 2 & -3\\0 & 0 & 2\end{bmatrix}$，$\rho(\boldsymbol{G})=2>1$，所以发散；

 (3) $\boldsymbol{G}=\begin{bmatrix}0 & 2 & -2\\0 & 2 & -1\\0 & 8 & -6\end{bmatrix}$，$\rho(\boldsymbol{G})=2+2\sqrt{2}>1$，所以发散；

(4) $G=\begin{bmatrix}0 & -0.5 & 0.5\\0 & -0.5 & 1.5\\0 & 0.5 & -1\end{bmatrix}$，$\rho(G)=\alpha\in(-1,-2)$，所以发散。

19. 把原方程组改写为

$$\begin{cases}10x_1-2x_2=3\\2x_1+10x_2-x_3=15\\x_1+2x_2-5x_3=10\end{cases}$$

由于此时的系数矩阵是主对角线元素按行严格占优阵，故按此形式使用Jacobi迭代法必收敛，迭代公式为

$$\begin{cases}x_1^{(k+1)}=0.2x_2^{(k)}+0.3\\x_2^{(k+1)}=-0.2x_1^{(k)}+0.1x_3^{(k)}+1.5\\x_3^{(k+1)}=0.2x_1^{(k)}+0.4x_2^{(k)}-2\end{cases}\quad(k=0,1,2\cdots)$$

20. $\rho(G_J)=\sqrt{\dfrac{a_{21}a_{12}}{a_{11}a_{22}}}$，$\rho(G_G)=\left|\dfrac{a_{21}a_{12}}{a_{11}a_{22}}\right|$。

由于$\rho(G_J)$和$\rho(G_G)$同时大于1，或同时等于1，也同时小于1，所以，Jacobi迭代法和Gauss-Seidel迭代法同时收敛或同时发散。

21. (1) 当且仅当$|a|<0.5$时，Jacobi迭代法收敛；

(2) 当且仅当$-0.5<a<1$时，Gauss-Seidel迭代法收敛。(实系数二次方程$x^2+px+q=0$的两个根之模均小于1的充分必要条件是：$|q|<1, 1+p+q>0, 1-p+q>0$)

22. 因A是正定矩阵，故用Gauss-Seidel迭代法求解必收敛。

23. 用$(0,0,0)^T$作初始向量进行迭代，得$x^*\approx x^{(10)}=(-3.999\ 997\ 829, 3.000\ 000\ 078, 1.999\ 999\ 583)^T$，迭代矩阵为

$$G_S=\begin{bmatrix}0.1 & -0.36 & -0.18\\0.022\ 5 & 0.019 & -0.490\ 5\\-0.011\ 925 & 0.069\ 93 & -0.000\ 035\end{bmatrix}$$

因为$\|G_S\|_\infty=0.54<1$，故迭代过程必收敛。

24. 用$(0,0,0)^T$作初始向量进行迭代，得

$$x^*\approx x^{(22)}=(0.500\ 004\ 397, 0.999\ 993\ 127, 0.999\ 997\ 183)^T$$

因方程组的系数矩阵A是正定矩阵，并且松弛因子$\omega=1.25$满足$0<\omega<2$，所以迭代过程必收敛。

第3章

1. $u_k=A^ku_0$，要用$(k-1)n^3+n^2$次乘法运算。$u_i=Au_{i-1}(i=1,2,\cdots,k)$，只须用$kn^2$次乘法运算。

2. (1) 用第一种幂法迭代格式(3.7)，并取$u_0=(1,0,0)^T$，则结果是

$$\lambda_1\approx\beta_8=11.000\ 28$$

$$x_1\approx y_7=(0.371\ 392\ 6, 0.743\ 056\ 1, 0.556\ 718\ 2)^T$$

(2) 用第一种幂法迭代格式(3.7)，并取$u_0=(1,0,0)^T$，则结果为

$$\lambda_1\approx\beta_9=-4.000\ 062$$

$$x_1 \approx y_8 = (-0.8164946, -0.4082504, -0.4082504)^T$$

3. 用类似 3.1.1 小节所用的方法证明。

4. 用反幂法迭代格式(3.11)，并取 $u_0=(1,0,0)^T$，则结果为

$$\lambda_3 \approx \beta_{20}^{-1} = -1.999725$$

$$x_3 \approx y_{19} = (-0.1826496, -0.3650610, 0.9128908)^T$$

5. 对矩阵 $A-5I$ 使用反幂法迭代，并取 $u_0=(1,0,0)^T$，则结果为

$$\lambda \approx \beta_4^{-1} + 5 = 5.124763$$

$$x \approx (0.4307502, 0.3507096, -0.9354931)^T$$

6. 用带原点平移的反幂法求 λ_1 和 λ_2，用带原点平移的幂法求 λ_n。

7.

$$\lambda_1 \approx 2.536526, \quad x_1 \approx (0.5314834, 0.4614733, 0.7103293)^T$$

$$\lambda_2 \approx -0.01664728, \quad x_2 \approx (-0.7212071, 0.6863493, 0.09372795)^T$$

$$\lambda_3 \approx 1.480121, \quad x_3 \approx (-0.444281, -0.5621094, 0.697601)^T$$

8. 先证明 $\det(\lambda_i I_{5\times5} - T) = 0$ 和 $\det(\lambda_j I_{5\times5} - T) = 0$，再用特征向量的定义证明$(a_1, a_2, a_3, 0, 0)^T$ 和$(0,0,0,b_1,b_2)^T$ 是矩阵 T 的分别相应于 λ_i 和 λ_j 的特征向量。

9. $$H = \begin{bmatrix} -0.61\dot{3} & -0.29\dot{3} & 0.7\dot{3} \\ -0.29\dot{3} & 0.94\dot{6} & 0.1\dot{3} \\ 0.7\dot{3} & 0.1\dot{3} & 0.\dot{6} \end{bmatrix}, \quad u = \begin{bmatrix} 2.4 \\ 1.8 \\ 0 \end{bmatrix}。$$

10. $$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.6666667 & -0.3333333 & 0.666667 \\ 0 & -0.3333333 & 0.9333333 & -0.1333333 \\ 0 & -0.6666667 & -0.1333333 & 0.7333333 \end{bmatrix};$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.9984104 & 0.05636188 \\ 0 & 0 & 0.05636188 & 0.9984104 \end{bmatrix};$$

$$A \backsim H_2 H_1 A H_1 H_2 = \begin{bmatrix} 1 & -3 & 0 & 0 \\ -3 & 2.222222 & 2.759943 & 0 \\ 0 & 2.759943 & -2.077975 & -1.016208 \\ 0 & 0 & -1.016208 & 1.855753 \end{bmatrix}。$$

11. $$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.5773503 & 0.5773503 & -0.5773503 \\ 0 & 0.5773503 & 0.7886751 & 0.2113249 \\ 0 & -0.5773503 & 0.2113249 & 0.7886751 \end{bmatrix};$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.3841196 & -0.9232834 \\ 0 & 0 & -0.9232834 & 0.3841195 \end{bmatrix};$$

$$A \backsim H_2H_1AH_1H_2=\begin{bmatrix} 2 & -1.154\,701 & -4.420\,268 & -2.264\,488 \\ -1.732\,051 & -0.333\,333\,3 & 1.473\,714 & -1.431\,917 \\ 0 & 3.091\,206 & 1.647\,287 & 0.046\,993\,74 \\ 0 & 0 & 2.356\,395 & 1.686\,047 \end{bmatrix}。$$

12. $\lambda_1=2$， $\lambda_2=3.732\,050\,808$， $\lambda_3=0.267\,949\,192$。

第4章

1. (1) 所求的最大正根 s 在区间(1,2)内，结果为 $s\approx x_9=1.532\,226\,563$；

   (2) 方程只有一个根，并且在区间(0,1)内，结果为 $s\approx x_9=0.510\,742\,187$。

2. (1) 方程只有一个根，且在区间(0,0.5)内，迭代公式采用

$$\begin{cases} x_0 \in (0,0.5) \\ x_{k+1} = \dfrac{1}{10}(2-\mathrm{e}^{x_k}) \quad (k=0,1,2,\cdots) \end{cases}$$

   必收敛(利用定理4.1证明)。取 $x_0=0.25$，则方程的根 $s\approx x_8=0.090\,525\,104$ 满足精度要求；

   (2) 最小正根在区间(0,2.5)内，采用以下迭代公式必收敛(利用定理4.1证明)：

$$\begin{cases} x_0 \in [1,2.5] \\ x_{k+1} = 1+\arctan x_k \quad (k=0,1,2,\cdots) \end{cases}$$

   取 $x_0=2$，得所求根 $s\approx x_8=2.132\,267\,568$；

   (3) 最小正根在区间 $\left(0,\dfrac{\pi}{2}\right)$ 内，采用以下迭代公式必收敛(利用定理4.1证明)：

$$\begin{cases} x_0 \in [0.5,1.5] \\ x_{k+1} = \arccos \mathrm{e}^{-x_k} \quad (k=0,1,2,\cdots) \end{cases}$$

   取 $x_0=1$，得所求根 $s\approx x_{11}=1.292\,695\,336$。

3. 利用定理4.3证明和判断。收敛速度是线性的。

4. 先证明 $s=\sqrt{a}$ 是方程 $x=\dfrac{x(x^2+3a)}{3x^2+a}$ 的根，再利用定理4.4证明序列 $\{x_k\}$ 收敛于 $\sqrt{a}$，且有三阶收敛速度。

5. 采用迭代公式

$$\begin{cases} y_k = \dfrac{1}{10}(2-\mathrm{e}^{x_k}), \quad z_k = \dfrac{1}{10}(2-\mathrm{e}^{y_k}) \\ x_{k+1} = x_k - \dfrac{(y_k-x_k)^2}{z_k-2y_k+x_k} \end{cases} \quad (k=0,1,2,\cdots)$$

   取 $x_0=0.25$，得方程的根 $s\approx x_3=0.090\,525\,101$。

6. (1) 取 $x_0=0.5$，得方程的根 $s\approx x_3=0.510\,973\,429$；

   (2) 此方程只有一个根，且在区间(0,3)内，取 $x_0=2$，结果是 $s\approx x_4=1.715\,620\,735$；

   (3) 所求根在区间$(31.5\pi,32.5\pi)$内，取 $x_0=102.1$，结果是 $s\approx x_6=102.091\,966\,5$。

7. 推广定理4.6的证明方法可知 $f(x)$ 应满足以下的条件：

   (1) 在包含根 s 的某个开区间内 $f'''(x)$ 连续且 $f'(x)\neq 0$；

   (2) $f''(s)=0$， $f'''(s)\neq 0$。

8. 利用定理 4.4 可知，选取 $h(x)=-\dfrac{f''(x)}{2f'(x)}$。

9. 用割线法，若选取 $x_{-1}=0, x_0=1$，则结果为 $s\approx x_5=0.510\ 973\ 429$；
   用单点割线法，若选取 $x_0=0.2, x_1=1$，则结果为 $s\approx x_8=0.510\ 973\ 428$。

10. 采用以下迭代公式

$$\begin{cases} x_1^{(k+1)} = 1.2 - e^{-2x_2^{(k)}} \\ x_2^{(k+1)} = \dfrac{1}{2}(1.97 - e^{-x_1^{(k)}}) \end{cases} \quad (k = 0,1,2,\cdots)$$

因 $\rho(\boldsymbol{G}'(\boldsymbol{x}^*))\leqslant e^{-0.75}<1$，根据定理 4.14 可知，只要 $(x_1^{(0)}, x_2^{(0)})^{\mathrm{T}}$ 充分接近 $\boldsymbol{x}^*$，此迭代公式就收敛于 $\boldsymbol{x}^*$。选 $x_1^{(0)}=x_2^{(0)}=1$，迭代结果为

$$x_1^*\approx x_1^{(6)}=0.998\ 414\ 093$$
$$x_2^*\approx x_2^{(6)}=0.800\ 832\ 393$$

11. 选取 $x^{(0)}=y^{(0)}=1.4$，得方程组的解为

$$x^*\approx x^{(3)}=1.433\ 534$$
$$y^*\approx y^{(3)}=1.472\ 349$$

第 5 章

1. 利用范德蒙德行列式证明。
2. 利用范德蒙德行列式证明。
3. 只须找出一个不等于零的三阶行列式

$$\begin{vmatrix} \sin x_i & \sin 2x_i & \sin 3x_i \\ \sin x_{i+1} & \sin 2x_{i+1} & \sin 3x_{i+1} \\ \sin x_{i+2} & \sin 2x_{i+2} & \sin 3x_{i+2} \end{vmatrix}$$

   其中 x_i, x_{i+1}, x_{i+2} 是点集 X 上的某三个点。

4. $p_4(x)=8.75(x-11)(x-12)(x-13)(x-14)-$
   $38.333\ 333\ 33(x-10)(x-12)(x-13)(x-14)+$
   $60(x-10)(x-11)(x-13)(x-14)-$
   $39.166\ 666\ 67(x-10)(x-11)(x-12)(x-14)+$
   $9.583\ 333\ 333(x-10)(x-11)(x-12)(x-13)$。

5. (1) 利用插值公式的余项式(5.9)证明；
   (2) 把 $(x_k-x)^m$ 展开，再利用(1)以及恒等式 $\sum\limits_{j=0}^{m}(-1)^j\dbinom{m}{j}\equiv 0$ 证明。

6. 利用插值公式的余项式(5.9)证明。

7. (1) $e^{-0.23}\approx 0.794\ 549\ 1$，$|e^{-0.23}-0.794\ 549|\leqslant 1.6\times 10^{-5}$；
   (2) $e^{-0.14}\approx 0.869\ 533\ 81$，$|e^{-0.14}-0.869\ 533\ 8|\leqslant 1.8\times 10^{-4}$。

8. $f(x)$ 的四次 Newton 插值多项式

$$\begin{aligned} P_4(x) = {} & 21 + 4(x+2) - 1.8(x+2)(x+1.5) + \\ & 0.4(x+2)(x+1.5)(x-0.5) - \\ & \frac{2}{35}(x+2)(x+1.5)(x-0.5)(x-1) \end{aligned}$$

9. 用数学归纳法证明。

10. $f[x_0,x_1,\cdots,x_n]=a_n, f[x_0,x_1,\cdots,x_k]=0\ (k>n)$。

11. 利用一阶差商定义证明。

12. (1) $f(x)\approx\frac{1}{h}[(x_{k+1}-x)f(x_k)+(x-x_k)f(x_{k+1})]$,

$$R_1(x)=\frac{f''(\xi)}{2!}(x-x_k)(x-x_{k+1}),\quad \xi\in(x_k,x_{k+1});$$

(2) $f(x)\approx\frac{1}{h^2}\sum_{i=k-1}^{k+1}\left(\prod_{\substack{j=k-1\\ j\neq i}}^{k+1}\frac{x-x_j}{i-j}\right)f(x_i)$,

$$R_2(x)=\frac{f'''(\xi)}{3!}\prod_{j=k-1}^{k+1}(x-x_j),\quad \xi\in(x_{k-1},x_{k+1});$$

(3) $f(x)\approx\frac{1}{h^3}\sum_{i=k-1}^{k+2}\left(\prod_{\substack{j=k-1\\ j\neq i}}^{k+2}\frac{x-x_j}{i-j}\right)f(x_i)$,

$$R_3(x)=\frac{f^{(4)}(\xi)}{4!}\prod_{j=k-1}^{k+2}(x-x_j),\quad \xi\in(x_{k-1},x_{k+2})。$$

13. $P_2(x)=21+4(x+2)-1.8(x+2)(x+1.5), f(-1)\approx P_2(-1)=24.1$;

$Q_2(x)=22-2(x-0.5)+0(x-0.5)(x-1), f(0.8)\approx Q_2(0.8)=21.4$;

$P_3(x)=23-0.5(x+1.5)-0.6(x+1.5)(x-0.5)+$

$0.2(x+1.5)(x-0.5)(x-1), f(0)\approx P_3(0)=22.85$。

14. 设插值节点为 x_i, x_i+h, x_i+2h,被插值点为 $x_i+t(0\leqslant t\leqslant 2h)$,则截断误差 $R(x_i+t)$的估计为 $|R(x_i+t)|\leqslant\frac{1}{6}\max\limits_{0\leqslant t\leqslant 2h}|t(t-h)(t-2h)|=\frac{\sqrt{3}}{27}h^3$,令 $\frac{\sqrt{3}}{27}h^3\leqslant 10^{-6}$,解得 $0\leqslant h\leqslant 0.024\,98$,所以节点的步长 h 最大不能超过 0.024 98。

15. 因 $|R(x)|\leqslant\frac{1}{6}\max\limits_{0\leqslant t\leqslant 0.8}|t(t-0.4)(t-0.8)|=0.004\,1<0.005$,而且 $f(x)$的近似值 $\tilde{u}$ 的第一位非零数字在小数点后第一位,故能保证 $\tilde{u}$ 有二位有效数字。

16. (1) $f(0.8, 1.25)\approx 16.45$;

(2) $f(0.8, 1.25)\approx 16.63$;

(3) $f(0.8, 1.25)\approx 16.873\,75$。

17. $H_4(x)=1+x-x(x-1)+\frac{1}{2}(x-1)^2x(x-2)$,

$$f(x)-H_4(x)=\frac{f^{(5)}(\xi)}{5!}x(x-1)^2(x-2)^2,\quad \xi\in(\min(x,0),\max(x,2))。$$

18. $H_2(x)=f(x_0)+f[x_0,x_1](x-x_0)+\frac{f'(x_0)-f[x_0,x_1]}{x_0-x_1}(x-x_0)(x-x_1)$,

其中 $f[x_0,x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0}$;

$$f(x)-H_2(x)=\frac{f'''(\xi)}{3!}(x-x_0)^2(x-x_1), \xi\in(\min(x,x_0,x_1),\max(x,x_0,x_1))。$$

19. $H_3(x)=1-0.619\,533\,333x+x(x-1.5)(0.107\,585\,184x-0.413\,022\,222)$;

$$\max_{0\leqslant x\leqslant 1.5}|f(x)-H_3(x)|\leqslant\frac{1}{4!}\max_{0\leqslant x\leqslant 1.5}|x^2(x-1.5)^2|=0.013\,183\,593。$$

20. 用半截单项式定义证明。

21. 只须证明$(x_+^m)^{(m-1)}=m!\,x_+,x\in(-\infty,\infty)$，以及$(x_+^m)^{(m)}=\begin{cases}m!, & x>0\\ 0, & x<0\end{cases}$。

22. (1) 是，$s(x)=-x^3-3x^2-x+2+2x_+^3,-1\leqslant x\leqslant 1$；

(2) 不是。

23. 用三次样条函数的定义验证。$s(x)=1-x^3+(x-1)_+^3-2(x-2)_+^3,0\leqslant x\leqslant 3$。

24. $$\Omega_3(2(x-1))=\begin{cases}0, & x\leqslant 0\\ \frac{4}{3}x^3, & 0<x\leqslant 0.5\\ -4x^3+8x^2-4x+\frac{2}{3}, & 0.5<x\leqslant 1\\ 4x^3-16x^2+20x-\frac{22}{3}, & 1<x\leqslant 1.5\\ -\frac{4}{3}x^3+8x^2-16x+\frac{32}{3}, & 1.5<x\leqslant 2\\ 0, & x>2\end{cases}$$

内节点有：$x_1=0,x_2=0.5,x_3=1,x_4=1.5,x_5=2$。

(图形略)

25. $$s(x)=-\frac{357}{270}\Omega_3(x+1)+\frac{10}{3}\Omega_3(x)+\frac{1\,617}{270}\Omega_3(x-1)+\frac{392}{45}\Omega_3(x-2)+\frac{43}{6}\Omega_3(x-3)+\frac{208}{45}\Omega_3(x-4),\quad 0\leqslant x\leqslant 3。$$

26. $M_0=-3.6,\quad M_1=1.2,\quad M_2=4.8,\quad M_3=-2.4$；

$$s(x)=\begin{cases}-0.6(2-x)^3+0.2(x-1)^3+8.6(2-x)+5.8(x-1), & 1\leqslant x\leqslant 2\\ 0.2(3-x)^3+0.8(x-2)^3+5.8(3-x)+4.2(x-2), & 2<x\leqslant 3\\ 0.8(4-x)^3-0.4(x-3)^3+4.2(4-x)+7.4(x-3), & 3<x\leqslant 4\end{cases}$$

27. $c_0=1.5,\quad c_1=-0.5,\quad c_2=0.5,\quad c_3=-0.5$。

28. $c_0=1.187\,5,\quad c_1=0.044\,194+0.150\,888\mathrm{i}$；

$c_2=-0.062\,5(1-\mathrm{i}),\quad c_3=-0.044\,194+0.025\,888\mathrm{i}$；

$c_4=-0.062\,5,\quad c_5=-0.044\,194-0.025\,888\mathrm{i}$；

$c_6=-0.062\,5(1+\mathrm{i}),\quad c_7=0.044\,194-0.150\,888\mathrm{i}$。

29. 只须证明 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$在区间$[a,b]$上线性无关。利用线性无关的定义证明。

30. 利用正交多项式的性质 1 和性质 2 证明。

31. 在式(5.73)中，利用 $g_k(x)=\frac{1}{a_k}\varphi_k(x)(k=0,1,\cdots)$即可推出所要的递推公式。

32. $\varphi_0(x)\equiv 1,\quad \varphi_1(x)=x-\frac{2}{3},\quad \varphi_2(x)=x^2-\frac{6}{5}x+\frac{3}{10}$。

33. 令 $x=\cos\theta$，再利用三角恒等式证明。

34. $p(x)=0.117\,187\,499+1.640\,625x^2-0.820\,312\,499x^4,\quad -1\leqslant x\leqslant 1$。

35. 可选 Legendre 多项式为基函数，即 $H_3=\mathrm{Span}\{L_0(x),L_1(x),L_2(x),L_3(x)\}$。

所求的三次最佳平方逼近多项式为

$$p_3(x)=\frac{3}{\pi}x+\left(\frac{7}{\pi}-\frac{105}{\pi^3}\right)\frac{1}{2}(5x^3-3x),\quad -1\leqslant x\leqslant 1$$

36. $p_2(x)=1.5x^2-0.6x+0.05,\quad 0\leqslant x\leqslant 1$。

37. $a=\frac{12}{\pi}\left(1-\frac{10}{\pi^2}\right),\quad b=-\frac{60}{\pi^2}\left(1-\frac{12}{\pi^2}\right),\quad c=\frac{60}{\pi^3}\left(1-\frac{12}{\pi^2}\right)$。

38. 利用内积性质以及最佳平方逼近的充分必要条件证明。

39. $\arccos x\approx\frac{\pi}{2}T_0(x)-\frac{4}{\pi}\left[T_1(x)+\frac{T_3(x)}{3^2}+\frac{T_5(x)}{5^2}+\frac{T_7(x)}{7^2}\right]=$

$$\frac{\pi}{2}-\frac{4}{\pi}\left(\frac{76}{105}x+\frac{1\,736}{2\,205}x^3-\frac{2\,016}{1\,225}x^5+\frac{64}{49}x^7\right),\quad -1\leqslant x\leqslant 1。$$

40. $s(x)=0.162\,022\,229\Omega_1(3x)+0.599\,715\,954\Omega_1(3x-1)+$

$$0.817\,171\,194\Omega_1(3x-2)+1.004\,203\,461\Omega_1(3x-3)=$$

$$\begin{cases}1.313\,081\,175x+0.162\,022\,229, & 0\leqslant x\leqslant\frac{1}{3}\\ 0.652\,365\,72x+0.382\,260\,714, & \frac{1}{3}<x\leqslant\frac{2}{3}\\ 0.561\,096\,801x+0.443\,106\,66, & \frac{2}{3}<x\leqslant 1\end{cases}。$$

41. $y=0.972\,58+0.050\,035\,1x^2$。

42. (1) $y(x)=2.007\,143+2.251\,429x$,

$$I_1=\sum_{i=1}^{7}[y_i-y(x_i)]^2=0.000\,514\,3;$$

(2) $y(x)=1.997\,619+2.251\,429x+0.038\,095x^2$,

$$I_2=\sum_{i=1}^{7}[y_i-y(x_i)]^2=0.000\,038\,07;$$

(3) $y(x)=1.997\,619+2.243\,651x+0.038\,095x^2+0.017\,778x^3$,

$$I_3=\sum_{i=1}^{7}[y_i-y(x_i)]^2=0.000\,021\,43。$$

$I_1>I_2>I_3$。

43. $s=ct^{\lambda}$ 两边取对数得 $\ln s=\ln c+\lambda\ln t$，令 $y=\ln s,a=\ln c$，$x=\ln t$，把原公式化为 $y=a+\lambda x$，再利用最小二乘法确定 a 和 λ，最后得 $s=4.393\,6t^{-0.110\,76}$。

44. $c_1=\frac{1}{4}(6+3\sqrt{2}),\quad c_2=0,\quad c_3=\frac{1}{4}(-6+3\sqrt{2})$。

45. $\varphi_0(x)\equiv 1,\quad \varphi_1(x)=x-5.5,\quad \varphi_2(x)=x^2-11x+22$。

46. $p(x,y)=9.824\,2+5.036\,9x-0.839\,98x^2-1.669\,2y^2+0.143\,03xy^2-0.748\,78x^2y^2$，拟合精度 $\sigma=41.007$（误差平方和）。

第6章

1. (1) 一次；(2) 三次。

2. $\lambda_0=\lambda_2=\frac{8}{3}h,\lambda_1=-\frac{4}{3}h$，有三次代数精度。

3. $x_0=-\frac{1}{\sqrt{3}},x_1=\frac{1}{\sqrt{3}}$，有三次代数精度。

4. $\int_{-1}^{1}f(x)\mathrm{d}x\approx\frac{1}{3h^2}[f(-h)+(6h^2-2)f(0)+f(h)]$。

截断误差 $R=\int_{-1}^{1}\frac{f'''(\xi)}{3!}x(x^2-h^2)\mathrm{d}x,\xi\in(-1,1)$ 且依赖于 x。

当 $h=\sqrt{0.6}$ 时，求积公式达到的最高代数精度为五次。

5. 只须证明在所给条件下求积公式(6.1)中的求积系数为 $\lambda_k=\int_a^b l_k(x)\mathrm{d}x\quad(k=0,1,\cdots,n)$，其中 $l_k(x)(k=0,1,\cdots,n)$ 是以求积节点 $x_0,x_1,\cdots,x_n$ 为插值节点的 Lagrange 插值基函数。为此，利用所给求积公式计算积分 $\int_a^b l_i(x)\mathrm{d}x\quad(i=0,1,\cdots,n)$ 的值即可证明。

6. (1) 设 $f(x)$ 在区间 $[a,b]$ 上有一阶连续导数，在区间 $[a,x]$ 上对 $f(x)$ 使用 Lagrange 中值定理即可推出求积公式的截断误差为 $R=\frac{f'(\eta)}{2}(b-a)^2,\eta\in(a,b)$。求积公式只有零次代数精度。

(2) 设 $f(x)$ 在区间 $[a,b]$ 上有二阶连续导数，利用 $f(x)$ 在点 $x_0=\frac{a+b}{2}$ 处展开的一阶 Taylor公式即可推出求积公式的截断误差为 $R=\frac{f''(\eta)}{24}(b-a)^3,\eta\in(a,b)$。求积公式具有一次代数精度。

7. 求积公式为 $\int_0^{3h}f(x)\mathrm{d}x\approx\frac{3}{4}h[f(0)+3f(2h)]$。

把 $f(x)$ 和 $f(2h)$ 在 $x_0=0$ 处分别展成 Taylor 级数即可证明求积公式的截断误差为

$$R=\frac{3}{8}h^4f'''(0)+O(h^5)$$

8. 用 $n=6$ 的复化梯形公式计算，得 $\int_1^2\frac{1}{x}\mathrm{e}^{-x}\mathrm{d}x\approx0.171\,945\,73$，$|R_{\mathrm{T}}|\leqslant0.004\,26$，所得近似值至少有二位有效数字。

用 $m=3$ 的复化 Simpson 公式计算，得 $\int_1^2\frac{1}{x}\mathrm{e}^{-x}\mathrm{d}x\approx0.170\,505\,777$，$|R_{\mathrm{S}}|\leqslant0.000\,102\,5$，所得近似值至少有三位有效数字。

9. 先判断该积分值的第一位非零数字在哪个数位上，然后确定复化梯形值的绝对误差限，最后利用复化梯形公式的截断误差确定 n 的值。结果为 n 至少取 58。

10. 利用复化梯形公式(6.12)的右端结构证明 $\frac{4T_{m+1}-T_m}{3}$ 就是 $2^{m+1}+1$ 个节点的复化 Simpson值[对照公式(6.14)的右端]。

11. $\int_0^1\mathrm{e}^{-x^2}\mathrm{d}x\approx T_6=0.746\,809\,1$。

12. $\int_0^1\mathrm{e}^{-x^2}\mathrm{d}x\approx T_1^{(3)}=0.746\,824\,1$。

13. 利用定理 6.1 即可说明其理由。
14. (1) 不属于; (2) 属于; (3) 属于。
15. (1) 1.568 627 451; (2)0 .746 824 466; (3) 0.170 482 66。
16. (1) 0.899 280 217;
  (2) 把 e^{-x^2} 写成 $e^{-x}e^{x-x^2}$,结果为 0.847 678 834;
  (3) 把$\frac{1}{1+x^2}$写成 $e^{-x}\frac{e^x}{1+x^2}$,结果为 1.376 569 829。
17. (1) 2.125 769 76; (2) 1.380 390 076。
18. (1) 2.052 344 305;
  (2) $a_0 \approx 3.977\ 462\ 634$, $a_1 \approx 1.775\ 494\ 647$, $a_2 \approx 0.426\ 393\ 23$。
19. (1) 使用 6.7.1 小节所讲的方法解题,结果是

$$x_1 = 0.112\ 008\ 806, \quad x_2 = 0.602\ 276\ 908$$
$$A_1 = 0.718\ 539\ 318, \quad A_2 = 0.281\ 460\ 82$$

  (2) 使用 6.7.1 小节所讲的方法解题,结果是

$$x_1 = 0.115\ 587\ 1, \quad x_2 = 0.741\ 555\ 747$$
$$A_1 = 1.304\ 290\ 305, \quad A_2 = 0.695\ 709\ 695$$

20. 只须验证三个求积节点是三次 Hermite 正交多项式的三个零点,并且三个求积系数符合相应的求积系数计算公式。
21. 若用 $n=3$ 的 Gauss - Legendre 求积公式计算,则结果为 0.632 591 389。
22. 积分约等于 18.6。
23. 积分约等于 7.167 176 973。

第 7 章

1. 求解结果:

t_n	0	0.1	0.2	0.3	0.4	0.5
y_n	1	0.9	0.828	0.776 001 6	0.739 063 8	0.714 004 8

2. 求解结果以及与精确解的比较:[其中 $e_n = y(t_n) - y_n$]

t_n	0	0.2	0.4	0.6	0.8
y_n	2	1.6	1.32	1.136	1.028 8
$y(t_n)$	2	1.656 192	1.410 960	1.246 435	1.147 987
e_n	0	0.056 192	0.090 960	0.110 435	0.119 187

3. 求解结果:

n	0	1	2	3	4	5
t_n	0	0.2	0.4	0.6	0.8	1.0
y_n	2	2.390 909	2.766 460	3.129 854	3.483 316	3.828 463
k_1	2	1.912 863	1.845 144	1.790 559	1.745 346	
k_2	1.909 091	1.842 644	1.788 803	1.744 057	1.706 119	

4. 求解结果以及与精确解比较：

n	0	1	2	3	4	5
t_n	0	0.2	0.4	0.6	0.8	1.0
y_n	2	2.438 178	2.956 333	3.554 473	4.232 603	4.990 726
k_1	2	2.398 481	2.797 379	3.196 542	3.595 888	
k_2	2.190 890	2.590 773	2.990 699	3.390 650	3.790 615	
$y(t_n)$	2	2.44	2.96	3.56	4.24	5.00
e_n	0	0.001 822	0.003 667	0.005 527	0.007 397	0.009 274

5. 求解结果：

n	0	1	2	3	4
t_n	0	0.2	0.4	0.6	0.8
y_n	2	1.656 2	1.410 973	1.246 450	1.148 004
k_1	−2	−1.456 2	−1.010 973	−0.646 450 5	
k_2	−1.7	−1.210 58	−0.809 875 5	−0.481 805 4	
k_3	−1.73	−1.235 142	−0.829 985 3	−0.498 269 9	
k_4	−1.454	−1.009 172	−0.644 975 8	−0.346 796 5	
$y(t_n)$	2	1.656 192	1.410 960	1.246 435	1.147 987
e_n	0	-8×10^{-6}	-13×10^{-6}	-15×10^{-6}	-17×10^{-6}

6. 只须验证它是二级二阶 Runge - Kutta 方法。

7. (1)不相容； (2) 不相容。

8. 利用定理 7.3 可证明题目的结论，关键是证明改进的 Euler 法的增量函数在区域

$$D=\{(t,y,h)\mid t_0\leqslant t\leqslant T,\ |y|<\infty,0\leqslant h\leqslant h_0\}$$

内连续且对变量 y 和 h 都满足 Lipschitz 条件。

9. (1)$0<h<0.4$； (2) $0<h<\dfrac{4}{3\sqrt{3}}$。

10. $0<h<\min\limits_{0<t\leqslant 2}\left(t+\dfrac{1}{t}\right)=2$。

11. 把四阶 R - K 方法(7.27)用于求解模型方程 $y'=\lambda y$ 的初值问题即可推出该方法的绝对稳定区域。

12. 求解初值问题(7.68)时，$0<h<0.556$；求解初值问题(7.69)时，$0<h<1.07$。

13. $\alpha=1,\beta=2$，局部截断误差 $R(t_{n+1})=\dfrac{1}{3}h^3y'''(t_{n-1})+O(h^4)$，是二阶方法。

14. $y_{n+3}+18y_{n+2}-9y_{n+1}-10y_n=h(3f_n+18f_{n+1}+9f_{n+2})$，

局部截断误差 $R(t_{n+3})=\dfrac{1}{20}h^6y^{(6)}(t_n)+O(h^7)$，是五阶方法。

15. $y_{n+3}+\dfrac{27}{11}(y_{n+2}-y_{n+1})-y_n=\dfrac{h}{11}(3f_n+27f_{n+1}+27f_{n+2}+3f_{n+3})$，

局部截断误差 $R(t_{n+3})=-\dfrac{537}{1\,232}h^7y^{(7)}(t_n)+O(h^8)$，是六阶方法。

16. $y_{n+3}-\frac{1}{11}(18y_{n+2}-9y_{n+1}+2y_n)=\frac{6}{11}hf_{n+3}$，局部截断误差 $R(t_{n+3})=-\frac{3}{22}h^4y^{(4)}(t_n)+O(h^5)$。

17. 利用相容性条件和定理 7.4 证明。

18. 利用定理 7.4 证明。

19. (1) 不相容，也不满足根条件；(2) 相容，但不满足根条件；(3) 相容，也满足根条件。

20. $e_m=c_{11}3^m+c_{21}2^m+c_{22}m2^m, m\in\{0,1,\cdots\}$（$c_{11},c_{21},c_{22}$ 是任意常数）。

21. (1) $\left|\mu+\frac{2}{3}\right|<\frac{2}{3}$；

(2) 由 $\mu=x+\mathrm{i}y$ 复平面上的闭曲线

$$\begin{cases}x(\theta)=(-10+15\cos\theta-6\cos 2\theta+\cos 3\theta)/[12(1+\sin^2\theta)]\\ y(\theta)=(35\sin\theta-10\sin 2\theta-\sin 3\theta)/[12(1+\sin^2\theta)]\end{cases}\quad 0\leqslant\theta\leqslant 2\pi$$

所围成的有界开区域。

22. (1) $-18<\mu<0$ 及 $\mu=\frac{24}{11}$；(2) 无绝对稳定区间；(3) $\mu<0,\mu>6$。

23.
$$\begin{cases}(y_0,z_0)^{\mathrm T}=(1,2)^{\mathrm T}\\ \begin{bmatrix}k_{11}\\k_{21}\end{bmatrix}=\begin{bmatrix}nh+y_n-z_n\\ nhy_nz_n\end{bmatrix}\\ \begin{bmatrix}k_{12}\\k_{22}\end{bmatrix}=\begin{bmatrix}\left(n+\frac{1}{2}\right)h+y_n+\frac{h}{2}k_{11}-\left(z_n+\frac{h}{2}k_{21}\right)\\ \left(n+\frac{1}{2}\right)h\left(y_n+\frac{h}{2}k_{11}\right)\left(z_n+\frac{h}{2}k_{21}\right)\end{bmatrix}\\ \begin{bmatrix}k_{13}\\k_{23}\end{bmatrix}=\begin{bmatrix}\left(n+\frac{1}{2}\right)h+y_n+\frac{h}{2}k_{12}-\left(z_n+\frac{h}{2}k_{22}\right)\\ \left(n+\frac{1}{2}\right)h\left(y_n+\frac{h}{2}k_{12}\right)\left(z_n+\frac{h}{2}k_{22}\right)\end{bmatrix}\\ \begin{bmatrix}k_{14}\\k_{24}\end{bmatrix}=\begin{bmatrix}(n+1)h+y_n+hk_{13}-(z_n+hk_{23})\\ (n+1)h(y_n+hk_{13})(z_n+hk_{23})\end{bmatrix}\\ \begin{bmatrix}y_{n+1}\\z_{n+1}\end{bmatrix}=\begin{bmatrix}y_n\\z_n\end{bmatrix}+\frac{h}{6}\begin{bmatrix}k_{11}+2k_{12}+2k_{13}+k_{14}\\k_{21}+2k_{22}+2k_{23}+k_{24}\end{bmatrix}\\ \qquad\left(n=0,1,\cdots,\left[\frac{10}{h}\right]-1\right)\end{cases}$$

24.
$$\begin{cases}(y_0,z_0)^{\mathrm T}=(1,1)^{\mathrm T}\\ (y_1,z_1)^{\mathrm T}\ \text{由与方法(7.70)同阶的单步法计算}\\ y_{n+2}=y_n+\frac{h}{2}(z_{n+1}+3z_n)\\ z_{n+2}=z_n+\frac{h}{2}[-200(y_{n+1}+3y_n)-20(z_{n+1}+3z_n)]\\ \qquad\left(n=0,1,\cdots,\left[\frac{10}{h}\right]-2\right)\end{cases}$$

绝对稳定性对步长 h 的限制为 $0<h<\frac{1}{15}$。

求解结果为$(t_n, y_n)\left(n=0,1,\cdots,\left[\frac{10}{h}\right]\right)$，其中$t_n=nh$。

25. 求解结果：

t_n	0	0.4	0.8	1.2	1.6	2.0
y_n	1	0.933 179 9	0.739 041 3	0.438 788 6	0.072 059 33	−0.305 876 6

26. 若用改进的 Euler 法求解，则对步长 h 的限制为 $0<h<0.335$。

若用 $k=6$ 的 Gear 方法求解，则由于本题对任何 $h>0$ 都有 $-18°47'<180°-\arg(h\lambda)<18°47'$，因而对步长 $h>0$ 无限制。

27.
$$\begin{cases}(x_0, y_0, z_0)^{\mathrm{T}}=(0,1,2)^{\mathrm{T}} \\ \begin{bmatrix}k_{11}\\k_{21}\\k_{31}\end{bmatrix}=\begin{bmatrix}-2 & nh & -\cos nh\\0 & -2 & \sin nh\\0 & \sin nh & -2\end{bmatrix}\begin{bmatrix}x_n\\y_n\\z_n\end{bmatrix}+\begin{bmatrix}0\\nh\\-n^2h^2\end{bmatrix} \\ \begin{bmatrix}k_{12}\\k_{22}\\k_{32}\end{bmatrix}=\begin{bmatrix}-2 & \left(n+\frac{1}{2}\right)h & -\cos\left(n+\frac{1}{2}\right)h\\0 & -2 & \sin\left(n+\frac{1}{2}\right)h\\0 & \sin\left(n+\frac{1}{2}\right)h & -2\end{bmatrix}\begin{bmatrix}x_n+\frac{h}{2}k_{11}\\y_n+\frac{h}{2}k_{21}\\z_n+\frac{h}{2}k_{31}\end{bmatrix}+\begin{bmatrix}0\\\left(n+\frac{1}{2}\right)h\\-\left(n+\frac{1}{2}\right)^2h^2\end{bmatrix} \\ \begin{bmatrix}x_{n+1}\\y_{n+1}\\z_{n+1}\end{bmatrix}=\begin{bmatrix}x_n\\y_n\\z_n\end{bmatrix}+h\begin{bmatrix}k_{21}\\k_{22}\\k_{32}\end{bmatrix} \\ \qquad\left(n=0,1,\cdots,\left[\frac{10}{h}\right]-1\right)\end{cases}$$

绝对稳定性对步长 h 的限制为 $0<h<\frac{2}{3}$。

28. $0<h<1$。

29. 刚性比 $r=20$，绝对稳定性对步长 h 的限制为 $0<h<0.139$。

30. 至少要计算 3.5×10^8 步。

第 8 章

1. 利用 Taylor 级数证明。

2. 求解结果：

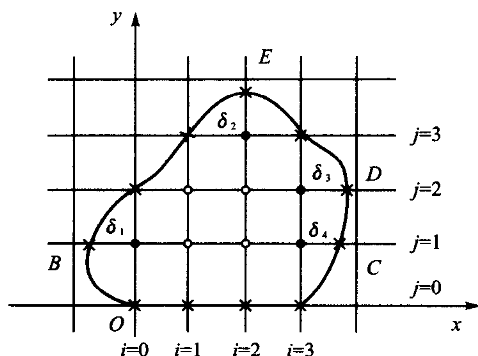
$h=1$ 的情形

u (y \ x)	−1	0	1
1	0	0	0
0	0	−4	0
−1	0	0	0

$h=0.5$ 的情形

u (y \ x)	−1	−0.5	0	0.5	1
1	0	0	0	0	0
0.5	0	−2.75	−3.50	−2.75	0
0	0	−3.50	−4.50	−3.50	0
−0.5	0	−2.75	−3.50	−2.75	0
−1	0	0	0	0	0

3. 正则内节点、非正则内节点和边界点如下图所示。



由上图得差分格式的矩阵形式：

$$
\begin{bmatrix}
1 & -\frac{\delta_1}{h+\delta_1} & & & & & & \\
1 & -4 & 1 & 1 & & & & \\
 & 1 & -4 & 0 & 1 & & & \\
 & 1 & 0 & -4 & 1 & 0 & 1 & \\
 & & 1 & 1 & -4 & 1 & 0 & 1 \\
 & & & 0 & -\frac{\delta_2}{h+\delta_2} & 1 & 0 & 0 \\
 & & & -\frac{\delta_4}{h+\delta_4} & 0 & 0 & 1 & 0 \\
 & & & & -\frac{\delta_3}{h+\delta_3} & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
u_{0,1} \\ u_{1,1} \\ u_{1,2} \\ u_{2,1} \\ u_{2,2} \\ u_{2,3} \\ u_{3,1} \\ u_{3,2}
\end{bmatrix}
=
\begin{bmatrix}
\frac{h}{h+\delta_1}\varphi(B) \\
h^2 f_{1,1}-\varphi(h,0) \\
h^2 f_{1,2}-\varphi(0,2h)-\varphi(h,3h) \\
h^2 f_{2,1}-\varphi(2h,0) \\
h^2 f_{2,2} \\
\frac{h}{h+\delta_2}\varphi(E) \\
\frac{h}{h+\delta_4}\varphi(C) \\
\frac{h}{h+\delta_3}\varphi(D)
\end{bmatrix}
$$

4. 求解结果：

y \ u \ x	0	1	2	3	4
3	0	0	0	0	0
2	2	0.878 1	0.432 5	0.214 7	0
1	2	1.080 0	0.734 8	0.426 4	0
0	0	0.707 1	1	0.707 1	0

5. 当 $\theta\neq-\frac{1}{2}$ 时，该差分格式对变量 t 是一阶精度的，对变量 x 是二阶精度的。当 $\theta=-\frac{1}{2}$ 时，该差分格式对变量 t 和对变量 x 都是二阶精度的。（利用 Taylor 级数判别）

6. (1) 求解结果：

t \ u \ x	0	0.2	0.4	0.6	0.8	1.0
$\frac{2}{150}$	0	0.542 22	0.853 33	0.853 33	0.542 22	0
$\frac{1}{150}$	0	0.586 67	0.906 67	0.906 67	0.586 67	0
0	0	0.64	0.96	0.96	0.64	0

(2) 求解结果：

t \ x (u)	0	0.2	0.4	0.6	0.8	1.0
0.2	−1.540 7	−1.848 9	2.693 3	2.693 3	−1.848 9	−1.540 7
0.1	0.977 78	1.173 3	0.16	0.16	1.173 3	0.977 78
0	0.533 33	0.64	0.96	0.96	0.64	0.533 33

7. 求解结果：

t \ x (u)	0	0.2	0.4	0.6	0.8	1.0
0.3	0	0.078 674 686	0.127 298 317	0.127 298 317	0.078 674 687	0
0.2	0	0.153 802 328	0.248 857 395	0.248 857 395	0.153 802 329	0
0.1	0	0.300 670 484	0.486 495 065	0.486 495 065	0.300 670 486	0
0	0	0.587 785 252	0.951 056 516	0.951 056 516	0.587 785 252	0

8. 齐次边界条件就是 $u_{0j}=u_{Nj}=0, j\geqslant 0$。用数学归纳法证明。

9. 无条件稳定。

10. 无条件稳定。

11. (1) 差分格式是 $u_{k,j+1}=u_{k,j-1}-ar(u_{k+1,j}-u_{k-1,j})$，其中 $r=\frac{\tau}{h}$。截断误差为 $O(\tau^2+h^2)$，是二阶精度的。差分格式的稳定性条件为 $r\leqslant\frac{1}{|a|}$。

(2) 差分格式是 $-aru_{k-1,j+1}+(1+ar)u_{k,j+1}=u_{kj}$，其中 $r=\frac{\tau}{h}$。截断误差为 $O(\tau+h)$，是一阶精度的。当 $a>0$ 时差分格式是无条件稳定的；当 $a<0$ 时差分格式的稳定性条件是 $r\geqslant\frac{1}{|a|}$。

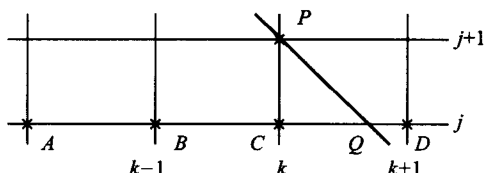
12. $u_{p_1}=-1.132\ 120\ 559$，$u_{p_2}=-2.877\ 756\ 598$，$u_{p_3}=-4.553\ 649\ 976$；
$u_{q_1}=-0.875$，$u_{q_2}=-2.457\ 031\ 25$，$u_{q_3}=-2.834\ 091\ 187$。

13. $u_{p_1}=8.4$，$u_{p_2}=19.44$，$u_{p_3}=38.832$。

14. 用类似于推导差分方程(8.72)的方法推导，有关的图形见下图，只须注意

$$u_Q=u_D+\frac{u_D-u_C}{x_{k+1}-x_k}(x_Q-x_{k+1})$$

$$x_Q=x_k+|CQ|=x_k-a\tau$$



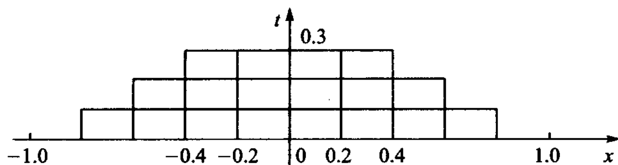
15. 求解结果(u 值只取三位有效数字)：

0.3	0.032 6	0.097 2	0.092 6	0.074 6	0.031 3	0
0.2	0.022 9	0.074 9	0.094 9	0.082 9	0.038 5	−0.020 0
0.1	0.012 0	0.070 0	0.096 0	0.090 0	0.052 0	−0.020 0
0	0	0.064 0	0.096 0	0.096 0	0.064 0	0
t / u / x	0	0.2	0.4	0.6	0.8	1.0

16. 计算格式为

$$\begin{cases}\begin{bmatrix}u_{k,j+1}^{(1)}\\u_{k,j+1}^{(2)}\end{bmatrix}=\begin{bmatrix}0.5 & -0.5\\-0.25 & 0.75\end{bmatrix}\begin{bmatrix}u_{k-1,j}^{(1)}\\u_{k-1,j}^{(2)}\end{bmatrix}+\begin{bmatrix}0.5 & 0.5\\0.25 & 0.25\end{bmatrix}\begin{bmatrix}u_{k+1,j}^{(1)}\\u_{k+1,j}^{(2)}\end{bmatrix}\\ \qquad [k=0,\pm1,\cdots,\pm(4-j);\ j=0,1,2]\\ \begin{bmatrix}u_{k,0}^{(1)}\\u_{k,0}^{(2)}\end{bmatrix}=\begin{bmatrix}0.2k(0.2k-1)\\0\end{bmatrix}\quad (k=0,\pm1,\pm2,\pm3,\pm4,\pm5)\end{cases}$$

所用的网格如下图所示。



求解结果：

t \ x	−1.0	−0.8	−0.6	−0.4	−0.2	0	0.2	0.4	0.6	0.8	1.0
0.3				$\begin{bmatrix}0.8\\-0.6\end{bmatrix}$	$\begin{bmatrix}0.48\\-0.48\end{bmatrix}$	$\begin{bmatrix}0.24\\-0.36\end{bmatrix}$	$\begin{bmatrix}0.08\\-0.24\end{bmatrix}$	$\begin{bmatrix}0\\0.12\end{bmatrix}$			
0.2			$\begin{bmatrix}1.08\\-0.46\end{bmatrix}$	$\begin{bmatrix}0.68\\-0.38\end{bmatrix}$	$\begin{bmatrix}0.36\\-0.3\end{bmatrix}$	$\begin{bmatrix}0.12\\-0.22\end{bmatrix}$	$\begin{bmatrix}-0.04\\-0.14\end{bmatrix}$	$\begin{bmatrix}-0.12\\-0.06\end{bmatrix}$	$\begin{bmatrix}-0.12\\0.06\end{bmatrix}$		
0.1		$\begin{bmatrix}1.48\\-0.26\end{bmatrix}$	$\begin{bmatrix}1\\-0.22\end{bmatrix}$	$\begin{bmatrix}0.6\\-0.18\end{bmatrix}$	$\begin{bmatrix}0.28\\-0.14\end{bmatrix}$	$\begin{bmatrix}0.04\\-0.1\end{bmatrix}$	$\begin{bmatrix}-0.12\\-0.06\end{bmatrix}$	$\begin{bmatrix}-0.2\\-0.02\end{bmatrix}$	$\begin{bmatrix}-0.2\\0.02\end{bmatrix}$	$\begin{bmatrix}-0.12\\0.06\end{bmatrix}$	
0	$\begin{bmatrix}2\\0\end{bmatrix}$	$\begin{bmatrix}1.44\\0\end{bmatrix}$	$\begin{bmatrix}0.96\\0\end{bmatrix}$	$\begin{bmatrix}0.56\\0\end{bmatrix}$	$\begin{bmatrix}0.24\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}-0.16\\0\end{bmatrix}$	$\begin{bmatrix}-0.24\\0\end{bmatrix}$	$\begin{bmatrix}-0.24\\0\end{bmatrix}$	$\begin{bmatrix}-0.16\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$

17. 用差分方程(8.72)求解的计算格式为

$$\begin{cases}\begin{bmatrix}u_{k,j+1}^{(1)}\\u_{k,j+1}^{(2)}\end{bmatrix}=\begin{bmatrix}1 & 1\\0.5 & 0.5\end{bmatrix}\begin{bmatrix}u_{k,j}^{(1)}\\u_{k,j}^{(2)}\end{bmatrix}+\begin{bmatrix}0 & -1\\-0.5 & 0.5\end{bmatrix}\begin{bmatrix}u_{k-1,j}^{(1)}\\u_{k-1,j}^{(2)}\end{bmatrix}\\ \qquad [k=0,1,2,-1,-2,\cdots,-(4-j);\ j=0,1,2]\\ \begin{bmatrix}u_{k,0}^{(1)}\\u_{k,0}^{(2)}\end{bmatrix}=\begin{bmatrix}0.2k(0.2k-1)\\0\end{bmatrix}\quad (k=0,1,2,-1,-2,-3,-4,-5)\end{cases}$$

用差分方程(8.73)求解的计算格式为

$$\begin{cases}\begin{bmatrix}u_{k,j+1}^{(1)}\\u_{k,j-1}^{(2)}\end{bmatrix}=\begin{bmatrix}1 & -1\\-0.5 & 1.5\end{bmatrix}\begin{bmatrix}u_{k,j}^{(1)}\\u_{k,j}^{(2)}\end{bmatrix}+\begin{bmatrix}0 & -1\\-0.5 & 0.5\end{bmatrix}\begin{bmatrix}u_{k+1,j}^{(1)}\\u_{k+1,j}^{(2)}\end{bmatrix}\\ \qquad [k=-1,-2,0,1,2,\cdots,(4-j);\ j=0,1,2]\\ \begin{bmatrix}u_{k,0}^{(1)}\\u_{k,0}^{(2)}\end{bmatrix}=\begin{bmatrix}0.2k(0.2k-1)\\0\end{bmatrix}\quad (k=0,-1,-2,1,2,3,4,5)\end{cases}$$

18. 该初值问题化为一阶双曲型方程组初值问题：

$$\begin{cases}\dfrac{\partial \boldsymbol{\omega}}{\partial t}+\boldsymbol{A}\dfrac{\partial \boldsymbol{\omega}}{\partial x}=\boldsymbol{0}, & 0<t<0.3,-\infty<x<\infty\\ \boldsymbol{\omega}(x,0)=(0,2x)^{\mathrm{T}}, & -\infty<x<\infty\end{cases}$$

其中 $\boldsymbol{\omega}=\begin{bmatrix}\omega^{(1)}\\\omega^{(2)}\end{bmatrix}$，$\boldsymbol{A}=\begin{bmatrix}0 & -1\\-1 & 0\end{bmatrix}$，$\omega^{(1)}=\dfrac{\partial u}{\partial t}$，$\omega^{(2)}=\dfrac{\partial u}{\partial x}$。求解初值问题(1)的计算格式为[使用格式(8.79)]

$$\begin{cases}\begin{bmatrix}\omega_{k,j+1}^{(1)}\\\omega_{k,j-1}^{(2)}\end{bmatrix}=\begin{bmatrix}0.5 & 0.5\\-0.25 & -0.25\end{bmatrix}\begin{bmatrix}\omega_{k-1,j}^{(1)}\\\omega_{k-1,j}^{(2)}\end{bmatrix}+\begin{bmatrix}0.5 & 0.25\\0.25 & 0.5\end{bmatrix}\begin{bmatrix}\omega_{k+1,j}^{(1)}\\\omega_{k+1,j}^{(2)}\end{bmatrix}\\ \qquad [k=0,\pm1,\pm2,\cdots,\pm(4-j);\ j=0,1,2]\\ (\omega_{k,0}^{(1)},\omega_{k,0}^{(2)})^{\mathrm{T}}=(0,0.4k)^{\mathrm{T}}\quad (k=0,\pm1,\pm2,\pm3,\pm4,\pm5)\end{cases}$$

ω 的计算结果见下表。

(j) t \ x ω (k)	−1.0 (−5)	−0.8 (−4)	−0.6 (−3)	−0.4 (−2)	−0.2 (−1)	0 (0)	0.2 (1)	0.4 (2)	0.6 (3)	0.8 (4)	1.0 (5)
(3) 0.3				$\begin{bmatrix}-0.46875\\0.26875\end{bmatrix}$	$\begin{bmatrix}-0.075\\0.275\end{bmatrix}$	$\begin{bmatrix}0.31875\\0.28125\end{bmatrix}$	$\begin{bmatrix}0.7125\\0.2875\end{bmatrix}$	$\begin{bmatrix}1.10625\\0.29735\end{bmatrix}$			
(2) 0.2			$\begin{bmatrix}-1.025\\0.225\end{bmatrix}$	$\begin{bmatrix}-0.65\\0.25\end{bmatrix}$	$\begin{bmatrix}-0.275\\0.275\end{bmatrix}$	$\begin{bmatrix}0.1\\0.3\end{bmatrix}$	$\begin{bmatrix}0.475\\0.325\end{bmatrix}$	$\begin{bmatrix}0.85\\0.375\end{bmatrix}$	$\begin{bmatrix}1.225\\0.375\end{bmatrix}$		
(1) 0.1		$\begin{bmatrix}-1.3\\-0.1\end{bmatrix}$	$\begin{bmatrix}-1\\0\end{bmatrix}$	$\begin{bmatrix}-0.7\\0.1\end{bmatrix}$	$\begin{bmatrix}-0.4\\0.2\end{bmatrix}$	$\begin{bmatrix}-0.1\\0.3\end{bmatrix}$	$\begin{bmatrix}0.2\\0.4\end{bmatrix}$	$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	$\begin{bmatrix}0.8\\0.6\end{bmatrix}$	$\begin{bmatrix}1.1\\0.7\end{bmatrix}$	
(0) 0	$\begin{bmatrix}0\\-2\end{bmatrix}$	$\begin{bmatrix}0\\-1.6\end{bmatrix}$	$\begin{bmatrix}0\\-1.2\end{bmatrix}$	$\begin{bmatrix}0\\-0.8\end{bmatrix}$	$\begin{bmatrix}0\\-0.4\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0.4\end{bmatrix}$	$\begin{bmatrix}0\\0.8\end{bmatrix}$	$\begin{bmatrix}0\\1.2\end{bmatrix}$	$\begin{bmatrix}0\\1.6\end{bmatrix}$	$\begin{bmatrix}0\\2\end{bmatrix}$

根据式(8.85)，可得

$$u_{kj}=1+(0.2k)^2+0.05\left(\omega_{k,0}^{(1)}+\omega_{k,j}^{(1)}+2\sum_{i=1}^{j-1}\omega_{k,i}^{(1)}\right)\quad (k=1,\pm1,\ \pm2;\ j=1,2,3)$$

由此可得本题的数值解为

t \ u x	−0.4	−0.2	0	0.2	0.4
0.3	1.001 562 5	0.968 75	1.015 937 5	1.143 125	1.350 312 5
0.2	1.057 5	0.986 25	0.995	1.083 75	1.252 5
0.1	1.125	1.02	0.995	1.05	1.185
0	1.16	1.04	1	1.04	1.16